




# Diversity and Complexity of the Large Surface Protein Family in the Compacted Genomes of Multiple *Pneumocystis* Species

 Liang Ma,<sup>a</sup> Zehua Chen,<sup>b,\*</sup> Da Wei Huang,<sup>c,\*</sup>  Ousmane H. Cissé,<sup>a</sup>  Jamie L. Rothenburger,<sup>d,\*</sup> Alice Latinne,<sup>e</sup> Lisa Bishop,<sup>a</sup> Robert Blair,<sup>f</sup> Jason M. Brechley,<sup>g</sup> Magali Chabé,<sup>h</sup> Xilong Deng,<sup>a</sup> Vanessa Hirsch,<sup>i</sup> Rebekah Keesler,<sup>j</sup> Geetha Kutty,<sup>a</sup> Yueqin Liu,<sup>a</sup> Daniel Margolis,<sup>a</sup> Serge Morand,<sup>k</sup> Bapi Pahar,<sup>f</sup> Li Peng,<sup>a</sup>  Koen K. A. Van Rompay,<sup>j</sup> Xiaohong Song,<sup>a</sup> Jun Song,<sup>l</sup> Antti Sukura,<sup>m</sup> Sabrina Thapar,<sup>a</sup> Honghui Wang,<sup>a</sup> Christiane Weissenbacher-Lang,<sup>n</sup> Jie Xu,<sup>l</sup> Chao-Hung Lee,<sup>o</sup> Claire Jardine,<sup>d</sup> Richard A. Lempicki,<sup>c</sup>  Melanie T. Cushion,<sup>p</sup>  Christina A. Cuomo,<sup>b</sup>  Joseph A. Kovacs<sup>a</sup>

<sup>a</sup>Critical Care Medicine Department, NIH Clinical Center, National Institutes of Health, Bethesda, Maryland, USA

<sup>b</sup>Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

<sup>c</sup>Leidos BioMedical Research, Inc., Frederick National Laboratory for Cancer Research, Frederick, Maryland, USA

<sup>d</sup>Department of Pathobiology, Canadian Wildlife Health Cooperative, Ontario Veterinary College, University of Guelph, Ontario, Canada

<sup>e</sup>EcoHealth Alliance, New York, New York, USA

<sup>f</sup>Tulane National Primate Research Center, Tulane University, New Orleans, Louisiana, USA

<sup>g</sup>Laboratory of Viral Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland, USA

<sup>h</sup>Université Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, U1019–UMR 8204–CILL–Centre d'Infection et d'Immunité de Lille, Lille, France

<sup>i</sup>Laboratory of Molecular Microbiology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland, USA

<sup>j</sup>California National Primate Research Center, University of California, Davis, Davis, California, USA

<sup>k</sup>Institut des Sciences de l'Évolution, Université de Montpellier 2, Montpellier, France

<sup>l</sup>Center for Advanced Models for Translational Sciences and Therapeutics, University of Michigan Medical Center, University of Michigan Medical School, Ann Arbor, Michigan, USA

<sup>m</sup>Department of Veterinary Pathology, Faculty of Veterinary Medicine, University of Helsinki, Helsinki, Finland

<sup>n</sup>Department of Pathobiology, Institute of Pathology, University of Veterinary Medicine Vienna, Vienna, Austria

<sup>o</sup>Department of Pathology and Laboratory Medicine, Indiana University School of Medicine, Indianapolis, Indiana, USA

<sup>p</sup>Department of Internal Medicine, College of Medicine, University of Cincinnati, Cincinnati, Ohio, USA

**ABSTRACT** *Pneumocystis*, a major opportunistic pathogen in patients with a broad range of immunodeficiencies, contains abundant surface proteins encoded by a multicopy gene family, termed the major surface glycoprotein (Msg) gene superfamily. This superfamily has been identified in all *Pneumocystis* species characterized to date, highlighting its important role in *Pneumocystis* biology. In this report, through a comprehensive and in-depth characterization of 459 *msg* genes from 7 *Pneumocystis* species, we demonstrate, for the first time, the phylogeny and evolution of conserved domains in Msg proteins and provide a detailed description of the classification, unique characteristics, and phylogenetic relatedness of five Msg families. We further describe, for the first time, the relative expression levels of individual *msg* families in two rodent *Pneumocystis* species, the substantial variability of the *msg* repertoires in *P. carinii* from laboratory and wild rats, and the distinct features of the expression site for the classic *msg* genes in *Pneumocystis* from 8 mammalian host species. Our analysis suggests multiple functions for this superfamily rather than just conferring antigenic variation to allow immune evasion as previously believed. This study provides a rich source of information that lays the foundation for the continued experimental exploration of the functions of the Msg superfamily in *Pneumocystis* biology.

**IMPORTANCE** *Pneumocystis* continues to be a major cause of disease in humans with immunodeficiency, especially those with HIV/AIDS and organ transplants, and is being seen with increasing frequency worldwide in patients treated with immunode-

**Citation** Ma L, Chen Z, Huang DW, Cissé OH, Rothenburger JL, Latinne A, Bishop L, Blair R, Brechley JM, Chabé M, Deng X, Hirsch V, Keesler R, Kutty G, Liu Y, Margolis D, Morand S, Pahar B, Peng L, Van Rompay KKA, Song X, Song J, Sukura A, Thapar S, Wang H, Weissenbacher-Lang C, Xu J, Lee C-H, Jardine C, Lempicki RA, Cushion MT, Cuomo CA, Kovacs JA. 2020. Diversity and complexity of the large surface protein family in the compacted genomes of multiple *Pneumocystis* species. *mBio* 11:e02878-19. <https://doi.org/10.1128/mBio.02878-19>.

**Editor** Louis M. Weiss, Albert Einstein College of Medicine

This is a work of the U.S. Government and is not subject to copyright protection in the United States. Foreign copyrights may apply.

Address correspondence to Liang Ma, [liang.ma@nih.gov](mailto:liang.ma@nih.gov).

\* Present address: Zehua Chen, ScholarIndex, Cambridge, Massachusetts, USA; Da Wei Huang, Laboratory of Molecular Biology, National Cancer Institute, Bethesda, Maryland, USA; Jamie L. Rothenburger, Department of Ecosystem and Public Health, Canadian Wildlife Health Cooperative (Alberta), Faculty of Veterinary Medicine, University of Calgary, Calgary, Alberta, Canada.

**Received** 31 October 2019

**Accepted** 16 January 2020

**Published** 3 March 2020

pleting monoclonal antibodies. Annual health care associated with *Pneumocystis* pneumonia costs ~\$475 million dollars in the United States alone. In addition to causing overt disease in immunodeficient individuals, *Pneumocystis* can cause sub-clinical infection or colonization in healthy individuals, which may play an important role in species preservation and disease transmission. Our work sheds new light on the diversity and complexity of the *msg* superfamily and strongly suggests that the versatility of this superfamily reflects multiple functions, including antigenic variation to allow immune evasion and optimal adaptation to host environmental conditions to promote efficient infection and transmission. These findings are essential to consider in developing new diagnostic and therapeutic strategies.

**KEYWORDS** classification, conserved domains, major surface glycoprotein, phylogenetic analysis, *Pneumocystis*

*Pneumocystis* continues to be a major cause of disease in humans with immunodeficiencies, especially those with HIV/AIDS and organ transplants, and is being seen with increasing frequency in patients treated with immunodepleting monoclonal antibodies. As an atypical fungus, *Pneumocystis* has highly adapted to the mammalian lung environment (1), with a high level of host specificity; *P. jirovecii* infects humans, *P. carinii* infects Norway rats (*Rattus norvegicus*), and *P. murina* infects house mice (*Mus musculus*). In addition, *Pneumocystis* cell walls are structurally unique and differ significantly from typical fungal cell walls that are composed of polysaccharides (mainly glucan and chitin) and highly mannosylated proteins. Both genomic and experimental analyses have shown the absence of chitin and outer chain *N*-mannans in *Pneumocystis* cell walls (1). Furthermore, beta-1,3-glucan is absent in the trophic form but masked in the cyst form of *Pneumocystis* (2).

An integral component of the *Pneumocystis* cell wall in both the cyst and trophic forms is the major surface glycoprotein (Msg) (also known as gp95, gp115, gp120, and gpA) (3–8). Ever since its identification in 1982 (9), Msg has been a focus of research, in part because it is the most abundant *Pneumocystis* protein as assessed by SDS-PAGE. Msg is present in all *Pneumocystis* species studied to date (3, 4, 6, 7, 10, 11) and appears to play an important role in pathogen-host interactions as well as in evasion of host immune responses. Based on studies of *Pneumocystis* in humans, rats, and mice, Msg is encoded by a multicopy gene family with an estimated ~30 to 100 copies per genome (5, 6, 8, 10, 12). *Msg* genes (up to ~3 kb each) are closely related to but clearly distinct from each other and are clustered together in the subtelomeric regions of multiple chromosomes (1, 13) (see Text S1 in the supplemental material). While there is no apparent variation in the *msg* repertoire among laboratory-bred *P. murina* or *P. carinii* isolates, extensive variation is present among *P. jirovecii* isolates (14).

Recently, we utilized long-read sequencing technology (15, 16) to identify the most complete set to date of *msg* genes in three *Pneumocystis* species (*P. jirovecii*, *P. carinii*, and *P. murina*) as part of the *Pneumocystis* genome project (1). Based on our studies, each *Pneumocystis* genome harbors approximately 60 to 180 *msg* genes, depending on the species, including the classical *msg* genes, *msg*-related (termed *msr*) genes, and additional related genes. These genes are collectively termed the *msg* superfamily. We previously reported on the first systematic classification of the *msg* superfamily (1) but did not provide a detailed description of the unique characteristics and phylogenetic relationships of individual domains and families or subfamilies. A recent report identified a small subset of *msg* genes in *P. jirovecii* from a single patient and described potential mechanisms of recombination, but this report did not include any other *Pneumocystis* species (17).

In the current report, we expanded our published analysis (1) to include *msg* genes in *Pneumocystis* from other mammalian host species. The goals of the current report were to (i) identify *msg* genes from *P. oryctolagi* (infecting rabbits), *P. wakefieldiae* (infecting rats), *Pneumocystis* sp. “*macacae*” (infecting rhesus macaques), and *Pneumocystis* sp. “*canis*” (infecting dogs); (ii) describe the characteristics and phylogenetic

evolution of individual *msg* domains; (iii) illustrate the characteristics, phylogenetic evolution, and relative expression levels of individual *msg* families or subfamilies; (iv) compare the variability of the *msg* repertoires in *P. carinii* from laboratory and wild Norway rats; and (v) characterize the variation of the expression sites or upstream conserved sequences (UCSs) of the classic *msg* genes in *Pneumocystis* from 11 mammalian host species.

## RESULTS

**Sources of *Pneumocystis msg* sequences.** *msg* sequences for *P. murina*, *P. carinii*, and *P. jirovecii* were obtained primarily from our previous *msg* and genome sequencing studies (1, 15, 16). The accuracy of these sequences was maximized by integrating Illumina high-throughput sequencing of genomic DNA, PacBio long-read sequencing of *msg* repertoire amplicons, and Sanger sequencing of cloned *msg* genes. Additional *msg* sequences were identified by Sanger sequencing of cloned *msg* amplicons and next-generation sequencing of whole genomes of the following *Pneumocystis* species: *P. wakefieldiae*, *P. carinii* (in wild rats), *P. oryctolagi*, *Pneumocystis* sp. "*macacae*," and *Pneumocystis* sp. "*canis*" (Table 1 and see Table S1 in the supplemental material). Due to the low-throughput nature and high cost of Sanger sequencing of cloned *msg* amplicons and the difficulty in assembling short reads from Illumina sequencing, only a small number of full-length *msg* genes were obtained from these species (1 to 13 genes per species). As whole-genome assembly of these species is still in progress, the *msg* genes reported for these species are only representative, not all inclusive. All *msg* sequences are available from the Zenodo database (data sets 1 to 8 available at <https://zenodo.org/record/3523554#.XbpSjld7mpo>) as well as the BioProject database with accession number (no.) PRJNA560924.

**Characteristics and phylogenetic relationships of individual Msg domains.** We identified a total of 9 conserved domains (Fig. 1 and Text S1). Classic Msg (Msg-A1) proteins contain 5 domains that presumably arose by gene duplication. Based on phylogenetic trees constructed using only these 5 domains, we found that each forms its own cluster, regardless of the origin of the species of the domains (Fig. 2). This strongly suggests that the most recent common ancestor to these *Pneumocystis* species already had developed this Msg domain structure and organization and that, subsequently, these domains evolved with no further duplication or recombination among domains across or within species. We also found that within each of these 5 Msg domains, individual domains clustered according to *Pneumocystis* species, suggesting that significant Msg family expansion occurred after the separation of *Pneumocystis* species. In addition, *P. carinii* and *P. murina* form two separate clusters, with each cluster containing both species, suggesting that those two clusters arose before separation of these two species.

The 31 N-linked glycans from *P. carinii* Msg proteins previously identified by liquid chromatography-tandem mass spectrometry (1) mapped to 4 domains, most commonly domains M4 and M5 (each with 13 glycans) and less commonly domains M2 (2 glycans) and M3 (3 glycans).

**Unique characteristics of each Msg family and subfamily.** Based on domain structure, phylogeny analysis, and expression control mechanisms of the *msg* superfamily, we previously proposed a classification of five families, named Msg-A, Msg-B, Msg-C, Msg-D, and Msg-E (1), as summarized in Table 1. According to the chromosome-level assemblies of the *P. murina*, *P. carinii*, and *P. jirovecii* genomes, *msg* genes are located almost exclusively in subtelomeric regions and are usually present in clusters (Text S1). Different *msg* families differ in the numbers of members, distributions among different *Pneumocystis* species, sequence structures (gene length, location and number of introns, and number of conserved domains), and expression control mechanisms, as summarized in Table 1. In addition, there is a bias of amino acid distribution among different Msg families (see Fig. S1 and Text S1).

**(i) Msg-A family.** Msg-A family is by far the largest among the 5 families of the Msg superfamily. This family is divided into three subfamilies: Msg-A1, Msg-A2, and Msg-A3

**TABLE 1** Summary of the *msg* superfamily members identified in *Pneumocystis* species

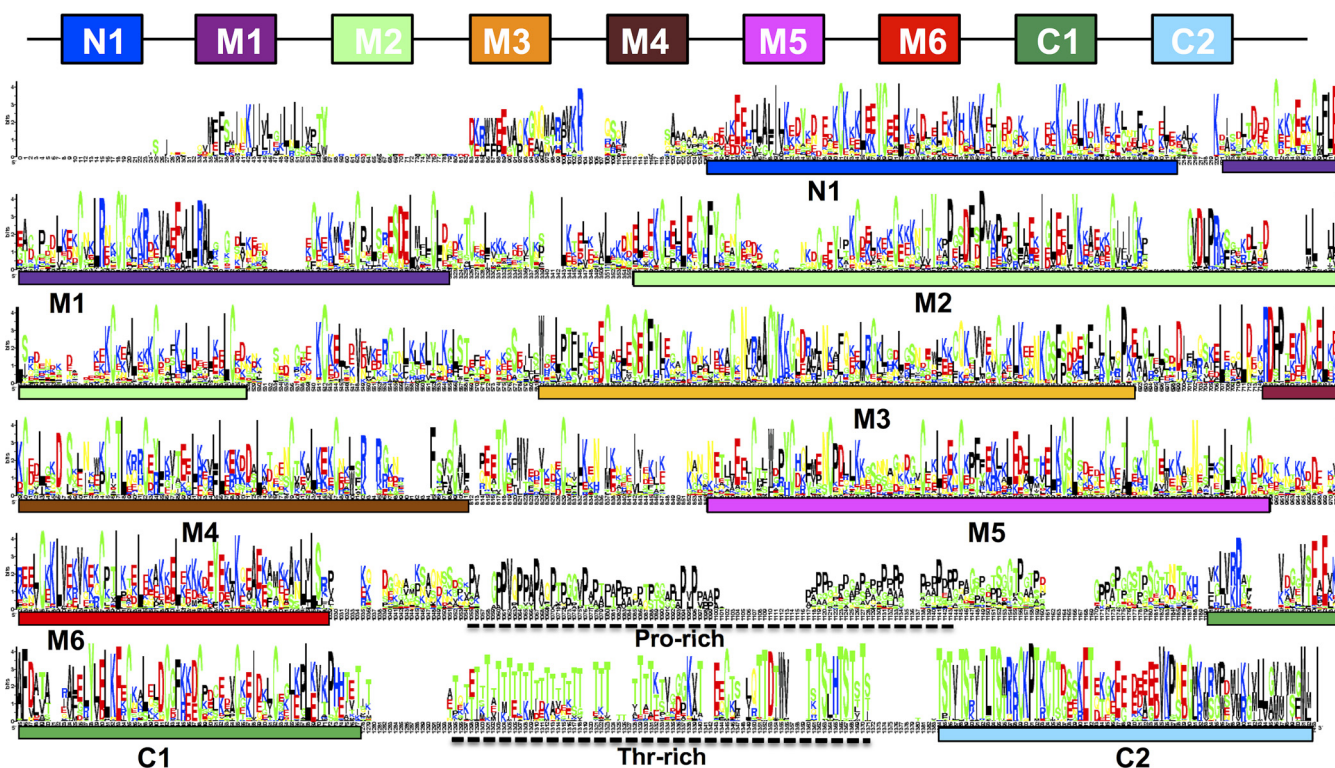
Family	No. of <i>msg</i> genes <sup>a</sup>										Characteristics <sup>b</sup>			
	<i>P. murina</i>	<i>P. carinii</i>	<i>P. jirovecii</i>	<i>P. wakefieldiae</i>	<i>P. sp. "macacae"</i>	<i>P. sp. "canis"</i>	<i>P. oryctolagi</i>	Avg gene (kb)/ protein (kDa) size	No. of introns	5'-end leader sequence	No. of domains	Expression mode		
<i>msg</i> -A1 <sup>c</sup>	22	65	86	2	2	9	7	3.2/120	0	CRJE <sup>d</sup>	9	Mutually exclusive		
<i>msg</i> -A2 <sup>c</sup>	14	53	0	6	0	0	0	2.9/218	1	Highly conserved	7-9	Independently		
<i>msg</i> -A3 <sup>c</sup>	6	3	33	2	3	1	0	3.1/117	1-2	Highly variable	9	Independently		
<i>msg</i> -B	0	0	21	0	2	0	1	1.4/55	1-2	Highly variable	2-3	Independently		
<i>msg</i> -C	6	1	2	2	1	0	0	1.7/60	1	Highly conserved	3	Independently		
<i>msg</i> -D	1	1	20	1	4	2	1	2.9/111	1	Highly variable	3-6	Independently		
<i>msg</i> -E	7	5	5	7	5	2	2	1.2/49	1	Highly variable	1	Independently		
Unclassified	8	13	12	0	0	0	0	0.9/37	0-8	Highly variable	0-1	Unknown		
Total	64	141	179	20	17	14	11							

<sup>a</sup>Results represent a potentially complete set for *P. murina*, *P. carinii*, and *P. jirovecii* as described in reference (1) but only a subset of members for the other four species (*P. wakefieldiae*, *P. oryctolagi*, *Pneumocystis* sp. "macacae," and *Pneumocystis* sp. "canis"), whose genome sequencing is ongoing.

<sup>b</sup>Based on *P. murina*, *P. carinii*, and *P. jirovecii* genes with complete or nearly complete reading frames.

<sup>c</sup>Subfamilies belonging to *msg*-A family.

<sup>d</sup>CRJE, conserved recombination joint element, which is unique, highly conserved among all *msg*-A1 genes.

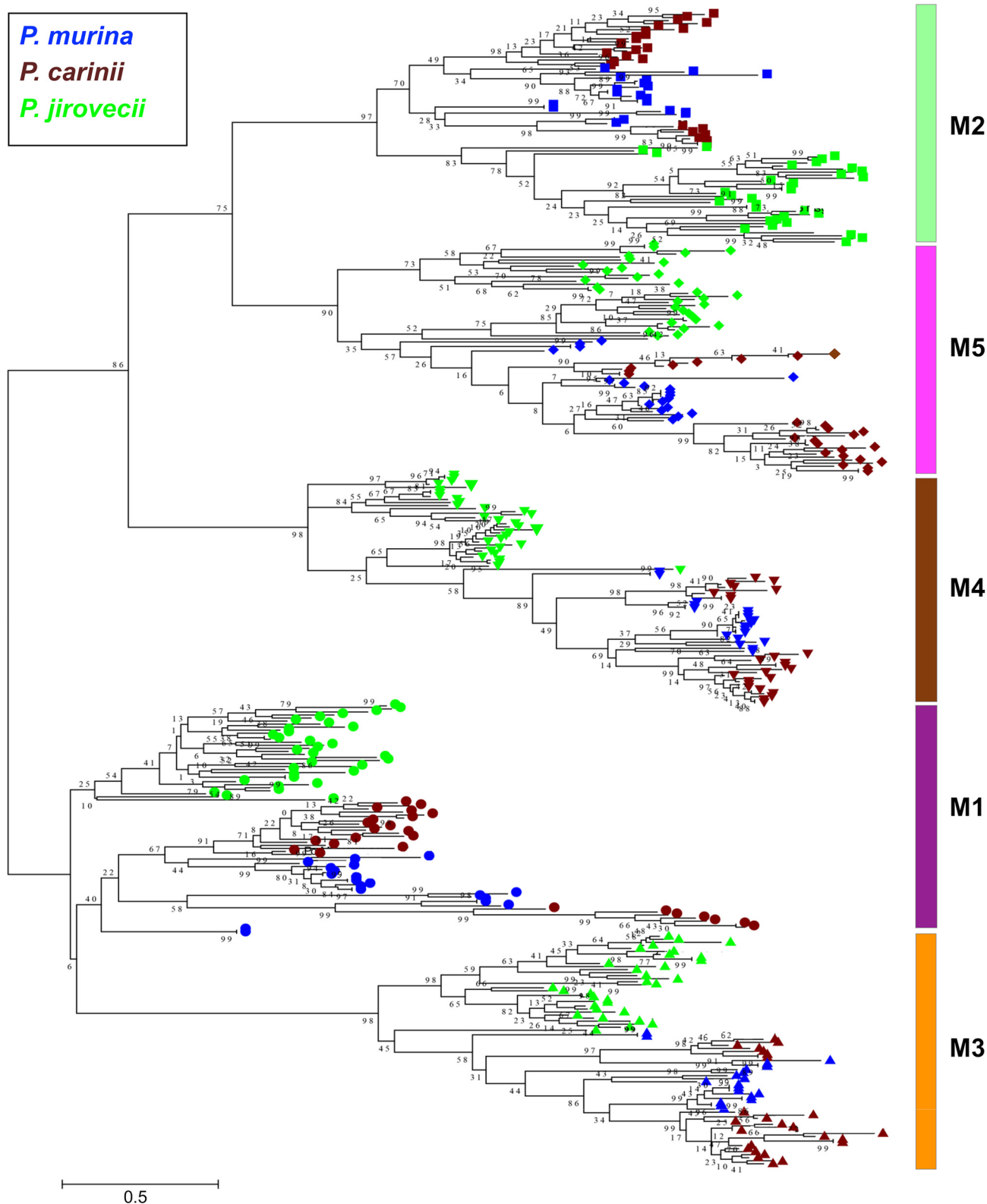


**FIG 1** Sequence logo showing the alignment of full-length Msg proteins in *P. murina*, *P. carinii*, and *P. jirovecii*. The known Pfam domains M1 to M5 (Pfam MSG) and C1 (extended from Pfam Msg2\_C to cover a longer conserved region), three new domains (N1, M6, and C2), and Pro- and Thr-rich regions are indicated. The horizontal axis represents the position of the amino acids. The vertical axis indicates conservation of each position as measured by information content (bits). This logo is adapted from Fig. 3 of reference 1, in which individual domains were shown separately without aligning with full-length proteins.

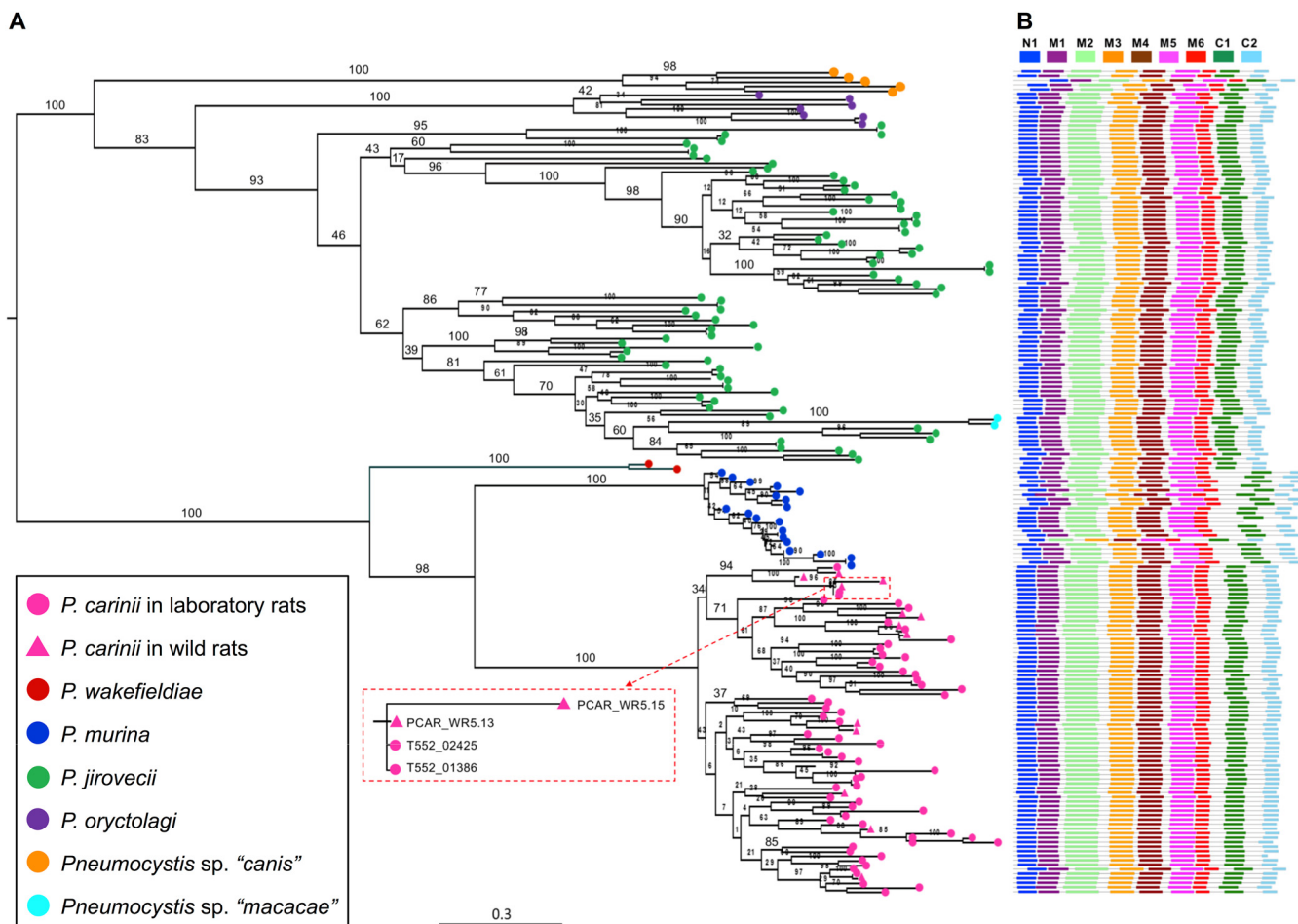
(Fig. 3 and Fig. S2 and S3). In addition to differences in phylogenetic relationships, these three subfamilies have significant differences in the expression control mechanisms and sequence structures of the 5'-end leaders.

The Msg-A1 subfamily includes all classic *msg* genes. The most striking characteristics of this subfamily are its dominance among all Msg families/subfamilies across all *Pneumocystis* species (Table 1) and its unique expression control mechanism. It has been well established that expression of this subfamily is controlled by a dedicated, single-copy subtelomeric expression site, also known as the upstream conserved sequence (UCS) (18–21) (Fig. 4). The UCS is expressed in fusion with an *msg* gene; the region between UCS and its downstream *msg* gene is termed the conserved recombination joint element (CRJE), which is highly conserved among all *msg*-A1 genes and potentially serves as an anchor for recombination (22). Available data suggest that these *msg* genes are not expressed unless they are translocated downstream of and in-frame with the UCS (18–21). This mechanism allows only one *msg*-A1 gene to be expressed in a single organism at a given time, although multiple *msg*-A1 genes are expressed at the population level in immunosuppressed hosts. In a phylogenetic analysis of 183 full-length *msg*-A1 genes from 7 *Pneumocystis* species, these genes clustered by species; as expected, genes from all three rodent *Pneumocystis* species formed a strong monophyletic group (Fig. 3). As previously noted (14), the Msg-A1 family in *P. jirovecii* is composed of two phylogenetically distinct groups; such separation is also seen in *P. murina* and *P. carinii*.

The Msg-A2 subfamily represents the previously reported *msr* genes in *P. carinii* (23, 24) and differs from Msg-A1 as follows: (i) the abundant presence in rodent *Pneumocystis*, but absence in all other species examined thus far (Table 1); (ii) the presence of a short intron near the 5' end; (iii) the presence of a unique highly conserved exon 1; and (iv) independent expression of each individual gene (not under the control of UCS).



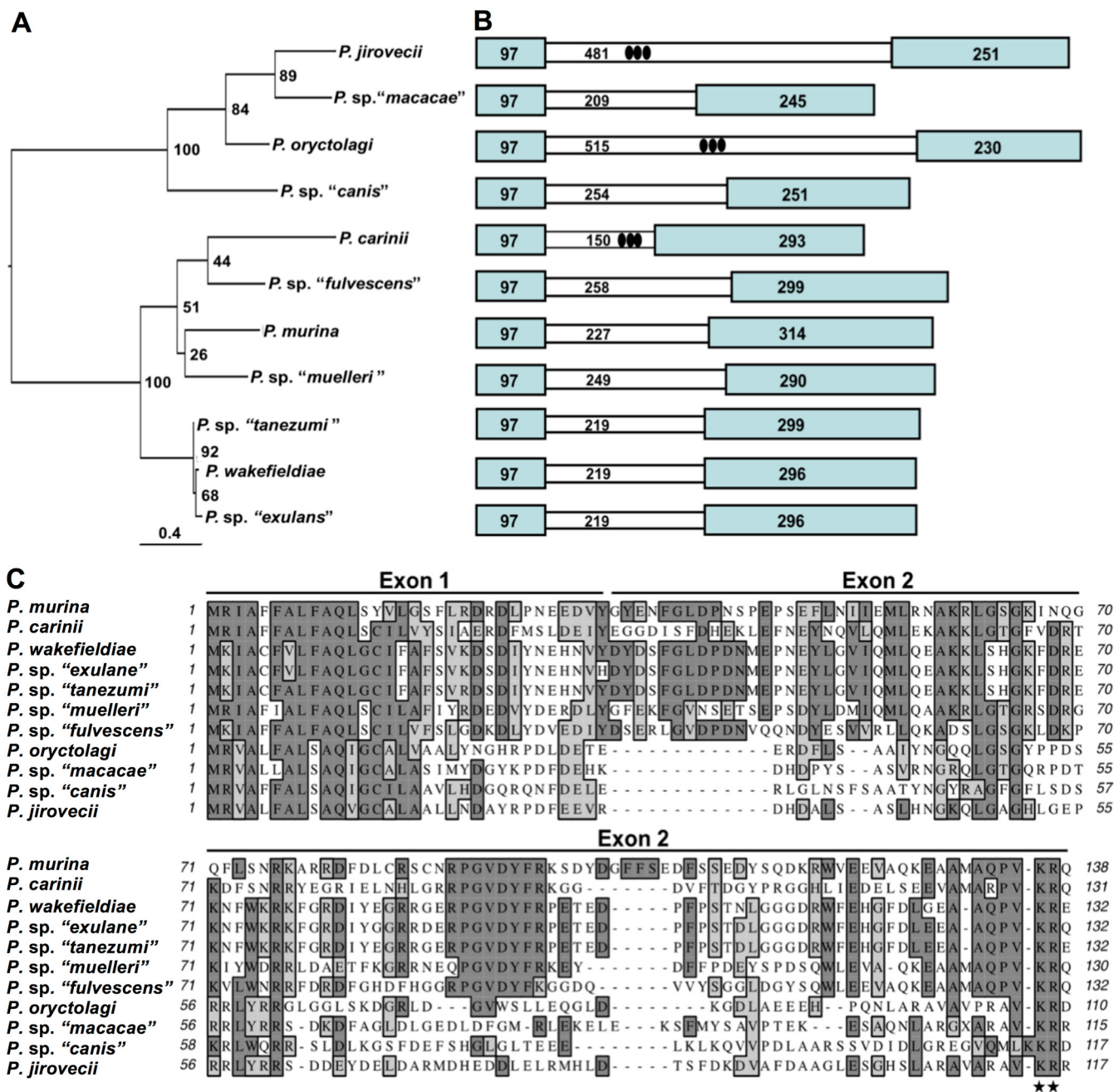
**FIG 2** Maximum likelihood tree based on aligned but not concatenated protein sequences of Pfam MSG domains M1 to M5 from Msg proteins longer than 900 amino acids in *P. murina*, *P. carinii*, and *P. jirovecii*. In the tree, different domains are indicated by different shapes on the right end of each branch, with different species color coded as shown at the top left. The color of each domain bar is the same as in Fig. 1. Numbers on the branch nodes indicate bootstrap support values.



**FIG 3** Phylogenetic tree and conserved domain structure of classic Msg genes (Msg-A1 subfamily). (A) A maximum likelihood (ML) tree constructed using deduced full-length protein sequences of *msg-A1* genes. Different *Pneumocystis* species are color coded as indicated at the bottom left. *P. carinii* isolates from laboratory rats and wild rats are indicated by pink dots and pink triangles, respectively. Only one of the 13 Msg proteins from the wild rat (PCAR\_WR5.13) was nearly identical (one amino acid difference) to two Msg proteins in *P. carinii* from laboratory rats (T552\_02425 and T552\_01386), as shown in the red box with a dashed line. Numbers on the branch nodes indicate bootstrap support values. All sequences shown are available from data set 1 at Zenodo database (<https://zenodo.org/record/3523554#.XjLZ7UBFyF4>). (B) Schematic representations of conserved Msg domains. Different domains are color coded as indicated at the top. Each row corresponds to the domain structure of the corresponding protein in panel A.

Alignment of 73 full-length *msg-A2* genes from rodent *Pneumocystis* revealed two groups of genes with sizes of ~2 kb and ~3 kb. The group with a larger size contains all 9 Msg domains, while the other group lacks three of them (M5, M6, and C1). In a phylogenetic analysis, all *msg-A2* genes from *P. carinii* formed a strong clade (with 99% bootstrap support), while *msg-A2* genes from *P. wakefieldiae* were interspersed among *msg-A2* genes from *P. murina* (Fig. S2A). In *P. carinii*, 11 *msg-A2* genes show higher sequence identities to *msg-A1* genes than other *msg-A2* genes (53% to 63% versus 35% to 44%) and are clustered together with *msg-A1* genes from *P. carinii* in a phylogenetic analysis (Fig. S2B). Similarly, one of the 6 *msg-A2* genes in *P. wakefieldiae* shows higher sequence identity to and is clustered with *msg-A1* genes.

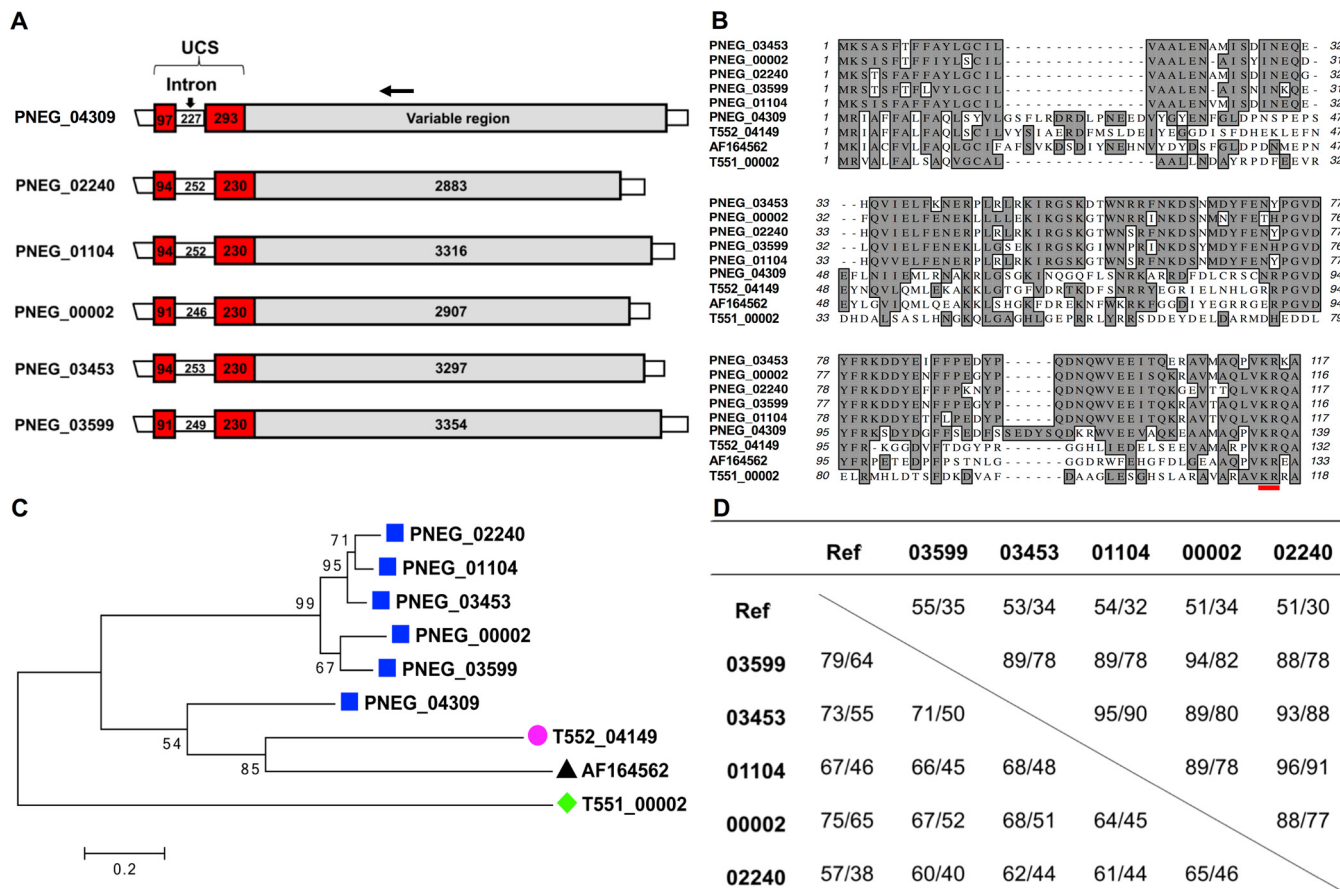
The Msg-A3 subfamily includes genes with substantial sequence identity to the Msg-A1 and Msg-A2 subfamilies but without the CRJE element of the *msg-A1* genes or the highly conserved exon 1 of the *msg-A2* genes (Fig. S3A). This subfamily has a significant expansion in *P. jirovecii* with 33 copies but only 1 to 6 copies in other species (Table 1). With an overall highly variable 5'-end leader, members of this subfamily are expected to be expressed independently, similarly to the Msg-A2 subfamily. Nevertheless, 5 of the 6 *msg-A3* genes in *P. murina* contain an ~600-bp leader sequence with significant identity and structural similarity to the UCS (termed UCS-like), including a



**FIG 4** Phylogeny and sequence structure of the expression sites or upstream conserved sequences (UCSs) of *msg-A1* genes in *Pneumocystis* species from different mammalian species. (A) Phylogenetic relationship based on protein sequences of UCSs. Numbers on the branches indicate bootstrap support values. (B) Schematic representation of the UCS sequence structures. The number in each box is the sequence length (base pairs) for each region. The approximate location of the tandem repeats in *P. jirovecii*, *P. oryctolagi*, and *P. carinii* are indicated by ovals, with more details on tandem repeat variation in *P. oryctolagi* and *P. carinii* shown in Fig. 6. (C) Alignment of deduced protein sequences of UCSs. Asterisks indicate the KR site potentially for proprotein cleavage by endoprotease. All sequences are available from GenBank with accession numbers or gene locus tag numbers indicated in parentheses, including *P. murina* (PNEG\_04309), *P. carinii* (T552\_04149), *P. wakefieldiae* (AF164562), *P. jirovecii* (T551\_00002), *P. oryctolagi* (MN509824), *Pneumocystis* sp. “*macacae*” (MN509821), *Pneumocystis* sp. “*canis*” (MN509823), *Pneumocystis* sp. “*fulvescens*” (MN509819), *Pneumocystis* sp. “*muelleri*” (MN509817), *Pneumocystis* sp. “*tanezumi*” (MN509820), and *Pneumocystis* sp. “*exulans*” (MN509818). Details about the nomenclature of *Pneumocystis* and its host species are available in Table S1 in the supplemental material.

relatively long intron (Fig. 5). These 5 genes are distributed in different chromosomes. Based on reverse transcription-PCR studies, each of these 5 genes was expressed independently (Text S1). Of note, the *Msg-A3* subfamily encompasses both the *Msg-II* and *Msg-III* families reported by others (17), as illustrated in Fig. S3B. Given the complex





**FIG 5** Five *msg-A3* genes containing a UCS-like leader sequence in *P. murina*. (A) Schematic representations of *msg* genes, including 5 containing a UCS-like sequence (PNEG\_02240, PNEG\_01104, PNEG\_00002, PNEG\_03453, and PNEG\_03599) and the UCS gene (PNEG\_04309) linked to one classical *msg-A1* gene (PNEG\_04308). The numbers in the boxes represent the size (base pairs) of the regions indicated. The horizontal arrow at the top indicates the approximate location of the reverse primer MSG.r2b conserved among all *msg-A1* and *msg-A3* genes and used to determine the expression of the 5 *msg-A3* genes (Table S2 and Text S1). (B) Alignment of the UCS and UCS-like protein sequences, including all those shown in panel A and the UCSs in *P. carinii* (T552\_04149), *P. wakefieldiae* (with GenBank accession no. AF164562), and *P. jirovecii* (T551\_00002). KR (red underlined) represents putative cleavage site by kexin-like endoprotease. (C) Phylogenetic relationship of UCS and UCS-like proteins based on sequences shown in panel B. Numbers on the branches indicate bootstrap support values. (D) Sequence identity among the *msg* genes shown in panel A (without including the first 4 characters of the gene identifiers [IDs]). Ref refers to the UCS gene PNEG\_04309 and the *msg-A1* gene PNEG\_04308 (linked downstream of PNEG\_04309). Values in the table refer to the identity (percent) of nucleotide and amino acid sequences in UCS (top right) and variable regions (lower left).

clustering patterns in phylogenetic analysis and unknown functions of these genes, it appears not meaningful to further divide the *Msg-A3* subfamily.

**(ii) Msg-B family.** This represents the only *Msg* family completely absent in all rodent *Pneumocystis* species sequenced to date but with exceptionally high abundance in *P. jirovecii* (Table 1). With a highly variable 5'-end leader, members of this family are expected to be expressed independently. In a phylogenetic analysis (see Fig. S4), the family separated into two major groups, which also differ in size (1.3 kb and 1.6 kb).

**(iii) Msg-C family.** The prominent characteristics of this family are its significant presence in *P. murina* and unique chromosomal organization (see Fig. S5). This family consists of a tandem array of 6 genes in chromosome 17 of *P. murina*, which represents the largest tandem array of genes of the same family identified so far in any *Pneumocystis* species. In contrast, there are no more than two *msg-C* genes in other *Pneumocystis* species. The two *msg-C* genes in *P. wakefieldiae* share similar sizes, intron-exon structures, and domain compositions (N1, M2, and M3) with *msg-C* genes in *P. murina*. However, in all other species examined, the *msg-C* genes are smaller (0.8 to 1 kb) with different intron-exon structures and/or lack the highly conserved exon 1 sequence of *P. murina*. In addition, they have different domain compositions and are only distantly related to the 6 genes in *P. murina* by phylogenetic analysis (Fig. S5A). Furthermore, the

**TABLE 2** Relative expression levels of the *msg* superfamily in *P. murina* and *P. carinii*

Genes	<i>P. murina</i>		<i>P. carinii</i>	
	No. of genes	FPKM <sup>a</sup>	No. of genes	FPKM <sup>a</sup>
<i>msg-A1</i>	22	670 (15–2,011)	65	141 (4–2,550)
<i>msg-A2</i>	14	252 (72–556)	53	205 (0–1,202)
<i>msg-A3</i>	6	725 (77–2,015)	3	112 (72–929)
<i>msg-C</i>	6	3,895 (662–9,382)	1	41
<i>msg-D</i>	1	1,385	1	2,154
<i>msg-E</i>	7	839 (100–4,270)	5	1,742 (122–8,010)
Unclassified	8	50 (0–90)	13	25 (8–158)
UCS	1	17,646	1	15,830
Genome	3,623	152 (0–17,646) <sup>b</sup>	3,646	148 (0–15,830) <sup>b</sup>

<sup>a</sup>FPKM, fragments per kilobase of exon per million fragments mapped based on RNA-Seq data as described in reference (1). Data are expressed as median (minimum to maximum) for multicopy gene families.

<sup>b</sup>Median values for all protein-coding genes in the genome.

chromosomal arrangement of the *msg-C* genes in *P. jirovecii* and *Pneumocystis* sp. “*macacae*” is different from that in rodent *Pneumocystis* (Fig. S5C). It is likely that the *msg-C* genes in *P. carinii*, *P. jirovecii*, and *Pneumocystis* sp. “*macacae*” represent degenerate genes or pseudogenes, as supported by the low-level transcription of this gene in *P. carinii* (Table 2).

**(iv) Msg-D family.** This family is related to the previously reported A12 antigen gene in *P. murina* (25). Like the Msg-A3 subfamily and Msg-B family, this family is rarely present in rodent *Pneumocystis* but significantly expanded in *P. jirovecii* and perhaps in *Pneumocystis* sp. “*macacae*” as well (Table 1 and Fig. S6). However, most of the Msg-D members contain 6 conserved domains compared to 9 and 3 domains in Msg-A3 and Msg-B, respectively. In a phylogenetic analysis, all single-copy Msg-D members in rodent *Pneumocystis* tightly clustered into one clade, which is well separated from Msg-D members in all other species. Consistent with the phylogenetic analysis, Msg-D members in all rodent *Pneumocystis* species lack both N1 and M2 domains, which are present in most of Msg-D members in other species.

**(v) Msg-E family.** This family is related to two previously reported p55 antigen genes (26–28). Unique among all Msg families, the Msg-E family is the smallest in member size, molecular size, and number of conserved domains. It is relatively equally distributed across all *Pneumocystis* species examined (Table 1) and among the most highly expressed families in rodent *Pneumocystis* (Table 2). Members did not cluster by species in a phylogenetic analysis (see Fig. S7). In *P. murina*, there are three members with nearly identical sequences and molecular sizes (termed p57), which are located in separate chromosomes (1, 29). Each of these 3 genes is present as a tandem array with one *msg-A2* gene and one *msg-A1* gene downstream (1). Homologs to these 3 genes are also present in three separate chromosomes of *P. wakefieldiae*, though their downstream genes have not been identified, presumably due to incomplete genome assembly. No other species sequenced to date have close homologs to these 3 genes. These findings further suggest that duplication of the p57 gene in *P. murina* and *P. wakefieldiae* occurred before separation of these two species or there was introgression.

**(vi) Unclassified genes.** In *P. murina*, *P. carinii*, and *P. jirovecii*, there are 8 to 13 genes related to Msg that are unable to be reliably classified due to their shorter length (~970 bp on average), presence of multiple introns, or lack of unique sequences (CRJE, KR site, or conserved leader sequences). The shorter length in most of them is not due to incomplete sequencing, as they are present within well-covered contigs. Almost all of these genes in *P. murina* and *P. carinii* have a low expression level (Table 2), suggesting they are degenerate genes or pseudogenes.

**Highly variable expression levels among different *msg* families in *P. murina* and *P. carinii*.** Transcriptome sequencing (RNA-Seq) data indicate that all *msg* genes in *P. murina* and *P. carinii* are transcribed except for two unclassified *msg* genes in *P. murina* and 5 *msg-A2* genes in *P. carinii* (Table 2). Strikingly, the UCS genes in both *P. murina* and *P. carinii* were the most highly expressed protein-coding genes of the

whole genome (with their fragments per kilobase of exon per million fragments mapped [FPKM] values being more than 100 times higher than the median expression level for the whole gene set); as expected, individual *msg-A1* members were expressed at lower levels. This high expression level is consistent with SDS-PAGE analysis of *Pneumocystis* proteins, which demonstrates that Msg is the most abundant protein as estimated by Coomassie blue staining (1). In *P. murina*, the highest expression level of individual Msg genes was observed in the Msg-C family, followed by the Msg-D, Msg-E, Msg-A3, and Msg-A1 families or subfamilies, all of which showed an expression level at least 3 times higher than the median of the whole gene set. The expression level of the Msg-A2 family was slightly higher than the median. In *P. carinii*, the highest expression level was observed in the single Msg-D gene, followed by the Msg-E family, both of which showed an expression level at least 11 times higher than the median. The expression level of the Msg-A family (including all 3 subfamilies) was similar to the median.

#### **Significant diversity of UCS in *Pneumocystis* from 10 mammalian host species.**

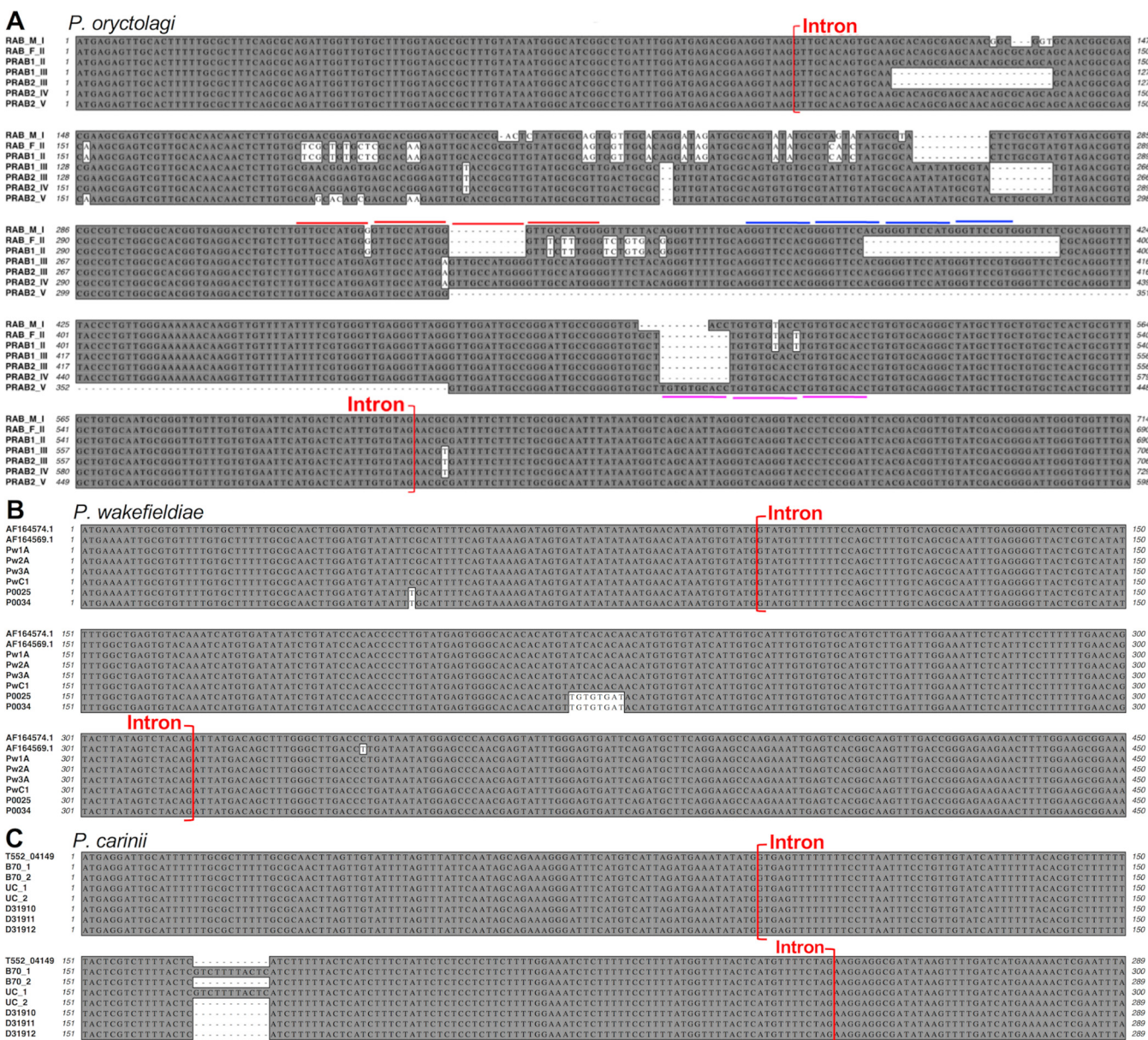
UCS was previously reported for *P. murina*, *P. carinii*, *P. wakefieldiae*, and *P. jirovecii* (18–21, 30). In the present study, we identified UCS in *Pneumocystis* species infecting rhesus macaques, dogs, rabbits, chestnut white-bellied rats, Müller's giant Sunda rats, Asian house rats, and Polynesian rats. Details about the nomenclature of these mammalian species and *Pneumocystis* species are listed in Table S1. As shown in Fig. 4, the *Pneumocystis* UCSs from all these animals show the sequence organization in known UCSs, including two exons that are interrupted by a variably sized intron. While exon 1 is identical in size (97 bp) among all UCSs, exon 2 is highly variable in size, with the shortest size present in *P. oryctolagi* (230 bp) and the longest in *P. murina* (314 bp).

The predicted UCS protein sequences vary in size from 110 to 138 amino acids, with 24% to 97% sequence identity (Fig. 4C). Despite these variations, all UCSs contain a pair of basic amino acid residues in the carboxyl end, Lys-Arg, known as the KR site (19, 31, 32). Phylogenetic analysis showed a clear separation between the UCSs in *Pneumocystis* species from rodents and those from other mammalian species (Fig. 4A). Consistent with the phylogenetic relationships, the UCSs in all rodent *Pneumocystis* species have an extra 13 to 15 amino acid residues at the beginning of exon 2 and a unique hexapeptide of PGVDYF near the center of exon 2 compared to *Pneumocystis* species from other mammalian species.

Similar to exon 2, the intron is also highly variable in size, with the shortest present in *P. carinii* (150 bp) and the longest in *P. oryctolagi* (515 bp). In addition, different levels of inter- or intrastain sequence variation were observed in UCSs from *P. carinii*, *P. wakefieldiae*, *Pneumocystis* sp. "*macacae*" and *P. oryctolagi* (Text S1 and Fig. 6). The highest variation was observed in *P. oryctolagi* isolates, which displayed extensive inter- and intrastain variations, including two single nucleotide polymorphisms (SNPs) in exon 2 and many SNPs, indels, and tandem repeat variations in the intron (Fig. 5).

**Substantial variation of the *msg-A1* gene repertoires in *P. carinii* from laboratory and wild rats.** To compare the *msg* diversity between *Pneumocystis* from laboratory-bred animals and that from wild animals, we analyzed the restriction fragment length polymorphism (RFLP) patterns of the *msg-A1* repertoires in *P. carinii* from 8 wild Norway rats collected in Ontario, Canada, in comparison with *P. carinii* from 8 laboratory Norway rats collected in three different animal facilities in United States. *P. carinii* from all laboratory rats showed almost identical RFLP patterns, while substantial variations in the RFLP patterns were observed within *P. carinii* isolates from wild rats and between *P. carinii* from wild and laboratory rats (Fig. 7).

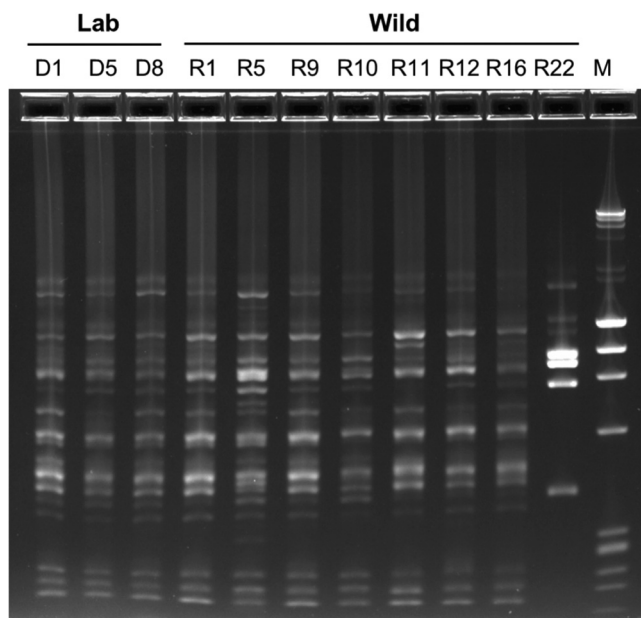
To further confirm these variations, we determined the full-length *msg-A1* sequences in *P. carinii* from one wild rat by Sanger sequencing of cloned PCR products. Sequence analysis of 28 random clones identified 13 unique *msg* sequences, with identities of 78% to 96% at the nucleotide level and 63% to 95% at the amino acid level. All but 1 of these 13 genes were clearly different from the 65 *msg-A1* genes of *P. carinii* from the laboratory rats. In a phylogenetic analysis, *msg-A1* genes from wild and laboratory rats were interspersed (Fig. 3). These findings suggest a closely related but clearly distinct repertoire of *msg-A1* genes in *P. carinii* from wild and laboratory rats.



**FIG 6** Genomic sequence variation in the expression site (UCS) of the *msg-A1* gene in *P. oryctolagi*, *P. wakefieldiae*, and *P. carinii*. (A) UCS in 4 *P. oryctolagi* isolates. Five different sequence populations were identified and named types I to V indicated at the end of the sample codes, including RAB\_M from MI, USA, RAB\_F from Tours, France, and PRAB1 and PRAB2 from Lille, France. The 3 types of sequences (III, IV, and V) in sample PRAB2 were obtained from sequencing of 2, 6, and 3 plasmid clones, respectively, while the 2 types of sequences (II and III) in sample PRAB1 were obtained from 3 and 5 plasmid clones, respectively. The other 2 samples showed no variation based on Illumina sequencing of genomic DNA; their PCR products showed homogeneous sequences in direct Sanger sequencing and were not further subcloned. Three types of tandem repeats are indicated by colored lines. (B) UCS in 8 *P. wakefieldiae* isolates. Sequences for the first two isolates were reported by Schaffzin et al. (24), with GenBank accession no. AF164574.1 and AF164569.1. Sequences for the other 6 isolates were obtained in this study, as determined by next-generation sequencing (NGS; isolates Pw1A, Pw2A, Pw3A, and PwC1 from laboratory rats) and PCR (isolates P0025 and P0034 from wild rats). (C) UCS in *P. carinii* isolates. The first sequence is the *P. carinii* UCS gene from the *P. carinii* genome assembly (1). The last 3 sequences indicated by GenBank accession no. D31910 to D31912 were reported by Wada et al. (21). Sequences B70\_1 and B70\_2 were assembled in this study using previous NGS data from one rat (1). Sequences UC\_1 and UC\_2 were assembled in this study using Sanger sequence reads from <http://pqp.cchmc.org> (73). The 11-bp tandem repeat unit is underlined. Numbers at both sides of the alignments refer to the nucleotide positions relative to the predicted UCS translational start site. The 3' end of exon 2 is not shown due to space limitation. The intron is indicated in red brackets. All new UCS sequences obtained in this study are available from GenBank with accession no. MN509813 to MN509830.

**DISCUSSION**

Over the past several decades, *Msg* has been the most extensively studied molecule in *Pneumocystis*, primarily due to its abundance, its role in pathogen-host interactions, and its potential as a target for diagnosis of *Pneumocystis* infection. In this report, we



**FIG 7** Comparative restriction fragment length polymorphism (RFLP) analysis of *msg-A1* of *P. carinii* infecting laboratory and wild Norway rats. *msg-A1* repertoire was amplified by PCR with genomic DNA prepared from *P. carinii*-infected lung samples from 8 immunosuppressed laboratory Norway rats (with three representatives show in lanes indicated Lab) and 8 wild Norway rats (in lanes indicated Wild). While R22 is most clearly representative of a different RFLP pattern, more subtle differences are also apparent in some of the other wild rats (e.g., R5 and R11) compared to that from laboratory rats. The PCR products were digested with restriction enzyme DraI and separated in 2% agarose gels containing SYBR Safe. Lane M, DNA size marker containing  $\lambda$ DNA digested with HindIII and  $\phi$ X174 DNA digested with HaeIII.

present an in-depth analysis of the *msg* domain structure and the characteristics of each individual *msg* family or subfamily, including new *msg* genes identified from *P. oryctolagi*, *Pneumocystis* sp. "*macacae*," *Pneumocystis* sp. "*canis*," *P. wakefieldiae*, and *P. carinii* infecting wild rats. The results from our analysis demonstrate a much greater complexity to this superfamily than was previously appreciated, expand the understanding of the primary structure, organization, phylogeny, and expression patterns of the Msg superfamily, and provide a comprehensive basis for further investigation of the role of the Msg superfamily in *Pneumocystis* biology.

The Msg superfamily, particularly, in *P. jirovecii* (179 members), represents the largest surface protein family identified to date in the fungal kingdom (33), which is surprising for an organism whose genome size is the smallest in the fungal kingdom sequenced to date, after the intracellular Microsporidia (34). *msg* genes are unique to *Pneumocystis* and account for 3% to 6% of the total genome, suggesting a critical role in the organism's survival (1). The vast majority of *msg* genes are clustered in subtelomeric regions, which are presumably advantageous to foster DNA recombination and antigenic variation, as has been found for surface protein genes in other pathogens (35). Their positioning is consistent with the notion that subtelomeric regions are favorable locations for fungal pathogens to acquire novel genes and foster evolution (36, 37).

By domain structure, phylogenetic relationships, and expression control mechanisms, we have been able to classify the Msg superfamily into discrete families and subfamilies. Our classification based on exhaustive cataloging of >400 full-length *msg* genes from seven *Pneumocystis* species is more comprehensive than the one described in a recent report, which was based on 113 *msg* genes from a single species, *P. jirovecii*, of which only 55 were full-length sequences (17). Thus, despite the consistency of four families/subfamilies between these two systems (Msg-A1, -B, -D, and -E versus Msg-I, -IV, -V, and -VI), two families (*msg-A2* and *msg-C*), which are almost exclusively present in rodent *Pneumocystis*, are absent in that report (17). We also elected not to subdivide the *msg-A3* subfamily due to the complex clustering patterns in the phylogenetic

analysis (see Fig. S4 in the supplemental material) and unknown functions of these genes. This classification will likely be refined when our understanding of the function of Msg is improved and this superfamily is better characterized for other *Pneumocystis* species.

Based on our analysis, there is substantial conservation among most Msg families or subfamilies across different *Pneumocystis* species, but there are also species-specific expansions or contractions. Among the three *Pneumocystis* species with the most complete data set, *P. murina* has the fewest number of genes in the Msg superfamily, while *P. jirovecii* has the most. These differences may be related to the larger body and therefore lung size, as well as the longer life span, in humans versus in rodents and the consequent need for a higher degree of antigenic variation to avoid the longer duration of immunologic memory in individual human hosts. The larger size of the Msg superfamily in *P. jirovecii* is attributable in part to the expansion of the classic Msg-A1 subfamily as well as other families (including Msg-A3, Msg-B, and Msg-D), which have no or limited representation in rodent *Pneumocystis*. Of note, *P. murina* possesses a set of 6 *msg* genes (Msg-C family) that are clustered as a tandem array in one chromosome and are the most highly expressed *msg* genes (Table 2).

The functions of Msg remain unknown or poorly understood. To date, the best studied genes of the Msg superfamily are those classical Msg genes in the Msg-A1 family, whose expression is regulated by the single-copy UCS expression site, which allows antigenic variation through DNA recombination (14, 17, 38). Such variation can potentially serve as a mechanism to facilitate evasion of host immune responses, enabling the organism to persist longer in the host and transmit to a new host. This mechanism evolved presumably to operate in immunocompetent hosts. The expression of multiple *msg*-A1 variants in the lungs of immunodeficient hosts presumably results from ongoing recombination at the UCS in the absence of immune pressure. For all three *Pneumocystis* species with nearly fully sequenced genomes, the *msg*-A1 genes account for approximately 50% of all *msg* genes, supporting their potential to efficiently generate a large number of variants allowing immune evasion. In support of this hypothesis, our RNA-Seq analysis of *P. murina* and *P. carinii* revealed an exceptionally high-level expression of UCSs and a variable level of expression of all individual *msg*-A1 genes (Table 2).

UCS is known to have a highly variable number of tandem repeats in the intron in *P. jirovecii* (19, 39, 40). In this study, we demonstrated for the first time the presence of inter- and intrastrain variations in tandem repeats in the intron of UCSs in *P. carinii* and *P. oryctolagi*. UCS in *P. oryctolagi* appears to have a higher degree of variation in tandem repeats as well as SNPs than *P. jirovecii* UCS. The intron in UCS (150 to 515 bp) is among the longest introns in *Pneumocystis* species studied to date. The retention of such a long intron with high variability in these species in an otherwise highly reduced genome suggests a critical role in organism survival, e.g., transcriptional regulation by a recursive splicing mechanism (41, 42).

Of note, while the UCS is present as a single-copy gene per genome in all *Pneumocystis* species, there are 5 *msg*-A3 genes in *P. murina*, each of which contains a UCS-like leader sequence (Fig. 5) and is expressed at a relatively high level independent of the classic UCS (Table 2). These may have arisen from gene duplication in *P. murina*; alternatively, it is possible that a common ancestor of *Pneumocystis* had multiple UCSs, which have been gradually lost as a result of evolving an efficient recombination system involving only a single UCS (for the *msg*-A1 family).

Previous studies have demonstrated a conservation of the *msg*-A1 repertoires in *Pneumocystis* in colony-bred laboratory rats and mice in contrast to the highly variable *msg* repertoires in *P. jirovecii* (14), suggesting a homogeneous population of rodent *Pneumocystis* due to closed breeding conditions. In support of this, we observed substantial variations in the RFLP patterns among *P. carinii* isolates from wild rats and between *P. carinii* from wild and laboratory rats, supporting the former possibility. The absence of clustering of Msg-A1 variants based on geographic origin of the isolates suggests that the repertoire variation was not driven by geographic isolation of the

organisms. These variations may reflect the difference in immune system development and modulation in wild animals, as they are continuously exposed to high levels of immune challenges in an open environment and experience high levels of infection with a wide range of pathogens (43–45). We hypothesize that this diversity is driven in part by a need for antigenic variation in response to T cell- rather than B cell-mediated immune responses and potential adaptation to the diverse HLA repertoire that would be present in a natural community of host species (46) versus the limited diversity present in inbred communities.

Domains M1 to M5 of *msg-A1*-encoded proteins likely arose by gene duplication given their conserved pattern of cysteine residues, and in fact, only a single M domain is categorized in Pfam. However, more detailed analysis clearly allows the identification of 5 related but unique domains. It is noteworthy that by phylogenetic analysis (Fig. 2), individual domains are observed as more closely related to each other across species than to other M domains in the same species, which suggests that there is a critical function for each domain and its evolution is restricted by negative selection. Furthermore, given that *msg-A1*-encoded proteins with these domains have been identified in all *Pneumocystis* species studied to date, this gene organization appears to have developed in an ancestor common to all *Pneumocystis* species and may have been a critical factor that allowed *Pneumocystis* to successfully infect mammalian hosts.

Unlike *P. jirovecii*, and perhaps *Pneumocystis* sp. “*macacae*,” *Pneumocystis* sp. “*canis*,” and *P. oryctolagi*, rodent *Pneumocystis* species (*P. murina*, *P. carinii*, and *P. wakefieldiae*) have a large number of *msg-A2* genes, which are only slightly less frequent than *msg-A1* genes. Previous studies of *P. carinii* have shown that *msg-A2* genes are expressed independent of the UCS (23, 24). Nevertheless, the possibility of homologous recombination between *msg-A2* and *msg-A1* genes cannot be ruled out due to their high sequence identities, as previously suggested (13, 47). Eleven *msg-A2* genes in *P. carinii* show higher identities to *msg-A1* genes than to other *msg-A2* genes in this organism. It is likely that these 11 *msg-A2* genes (the second exons) have arisen as a result of reciprocal recombination with *msg-A1* genes (through a mechanism unrelated to UCS or CRJE). While it appears that *msg-A2* expression is not regulated by UCS, nothing is known yet about what mechanisms control *msg-A2* expression or whether the *msg-A2* family contributes to antigenic variation in response to immune pressure, environmental changes, or life cycle phases. The presence of a long monoguanosine repeat in some *msg-A2* genes has raised the possibility that variation in the length of this repeat may cause frameshifts, thus altering the amino sequence downstream of the repeat (13, 47). However, based on the high-throughput genome sequencing data with at least 150× coverage (1), sequence reads for the monoguanosine repeat region in all involved *msg-A2* genes appeared highly uniform, though a small number of reads (<5%) showed different numbers of repeats. We could not determine if this was caused by sequence errors or *in vivo* changes. The presence of such a small number of variable reads does not seem to support an involvement of this repeat in altering the antigenicity or other functions of the *msg-A2* genes. Of note, a polyguanosine repeat encodes a polyglycine peptide, which has been shown in other organisms to play various critical roles, such as in protein-to-protein interactions, cell wall plasticity, and modulation of developmental stages (48–50). Whether the polyglycine peptide in *Msg-A2* proteins has these functions awaits future investigation.

Despite their potential importance in *Pneumocystis*’ survival, the functions of the vast majority of members of the *msg* superfamily remain poorly understood or uncharacterized. Even for the most extensively studied *msg-A1* genes, while it has been generally believed that their primary function is to confer antigenic variation and immune evasion, there are only limited experimental data supporting this potential function (46). There are also multiple studies showing an involvement of *Msg* proteins in mediating adherence of *Pneumocystis* organisms to host alveolar epithelial cells and macrophages (51–53), though it is unknown if the *Msg* proteins involved in these studies represent *Msg-A1* or other *Msg* proteins, especially *Msg-A2* and *Msg-A3* proteins, which are highly similar to *Msg-A1* proteins in sequence and length. The

functions of all non-*msg*-A1 genes remains unknown. Given that the *Pneumocystis* genome is highly compacted and that the DNA recombination system associated with *msg*-A1 genes is presumably sufficient for antigenic variation and immune evasion, there seems no reason to assume other *msg* genes perform the same function. We speculate that non-*msg*-A1 genes may perform other functions, such as mediating developmental states, optimizing mobility and adhesion ability, and adapting to specific host niches or environmental conditions. In support of this hypothesis, one such gene of the *msg*-E family in *P. murina*, termed p57, has been shown to be a stage-specific antigen that is expressed exclusively on intracystic bodies and young trophic forms, suggesting a role in the *Pneumocystis* life cycle development (29).

In conclusion, despite a highly reduced genome, *Pneumocystis* is equipped with a large complex superfamily of *msg* genes. These genes exhibit conservation among *msg* families and subfamilies across different *Pneumocystis* species as well as species-specific expansions or contractions. The versatility of these genes may mirror their association with a wide variety of functions rather than just conferring antigenic variation to allow immune evasion as previously believed. Our results provide a rich source of information that lays the foundation for the continued experimental exploration of the function of the *Msg* superfamily in *Pneumocystis* biology.

## MATERIALS AND METHODS

**Sources of *Pneumocystis msg* sequences.** The primary source of *msg* sequences for *P. murina*, *P. carinii*, and *P. jirovecii* was from our previous studies (1, 15, 16), which are available from the NCBI Umbrella project PRJNA223519 at <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA223519>. In this study, we obtained additional *Pneumocystis msg* and UCS sequences from various animals as listed in Table S1 in the supplemental material, which includes new tentative names for *Pneumocystis* organisms not reported previously. The methods to obtain these new sequences are described below.

***Pneumocystis* sample sources and DNA extraction.** Agarose gel blocks containing *P. wakefieldiae* and *P. carinii* were obtained from 4 Norway rats immunosuppressed once per week by 4 mg/kg of methylprednisolone acetate (Depo-Medrol; Pharmacia and Upjohn Co. a division of Pfizer, Inc.) at the animal facility of the University of Cincinnati, OH, USA (54). Genomic DNA in gel blocks was extracted using the ZymoClean Gel DNA Recovery kit (Zymo Research).

*P. carinii*-infected lung tissues were obtained from 8 immunosuppressed Sprague Dawley rats collected between 1986 and 2013 from the animal facilities at NIH, Bethesda, MD (14, 55), Indiana University, Indianapolis, IN (56), and Louisiana State University Health Science Center, New Orleans, LA. Genomic DNA was isolated by use of either a QIAamp DNA minikit (Qiagen) or a traditional method utilizing proteinase K digestion, phenol-chloroform extraction, and ethanol precipitation (14). In addition, *P. carinii*-infected lung tissues were obtained from 8 wild Norway rats (*R. norvegicus*) from 5 different pig farms in Ontario, Canada, in 2015 as previously described (57); exact locations and names of these farms were kept anonymous based on agreement with farm owners. Genomic DNA was extracted using the MasterPure Yeast DNA purification kit (Epicentre). All wild rats appeared to be healthy upon capture and were confirmed to be infected by *P. carinii* alone based on PCR and sequence analysis of two *Pneumocystis* mitochondrial genes, including the large subunit of rRNA (mtLSU) and ATPase subunit 6 genes (unpublished data). The *P. carinii* mtLSU sequence in all rats was identical to that in the laboratory Norway rats (GenBank JX499145).

DNA samples for *Pneumocystis* species infecting other wild rat species in Southeast Asia, including chestnut white-bellied rats, Müller's giant Sunda rats, Asian house rats, and Polynesian rats, were obtained from previous studies (58). All animals appeared to be healthy upon capture.

*Pneumocystis* sp. "*macacae*"-infected lungs were obtained from two simian immunodeficiency virus (SIV)-infected rhesus macaques at the NIH Animal Center, Bethesda, MD, USA (59, 60). Genomic DNA was extracted following a *Pneumocystis* DNA enrichment protocol as described previously (1). An additional two *Pneumocystis* sp. "*macacae*" samples were obtained as formalin-fixed paraffin-embedded (FFPE) tissue sections prepared from SIV-infected rhesus macaques at the Tulane National Primate Research Center, Covington, LA (61), and the UC Davis California National Primate Research Center, Davis, CA, USA. Genomic DNA was extracted using the AIIPrep DNA/RNA FFPE kit (Qiagen).

*Pneumocystis* sp. "*canis*" DNA samples were obtained from one Cavalier King Charles Spaniel dog at the University of Helsinki, Finland (62, 63), and one Whippet dog at the University of Veterinary Medicine, Vienna, Austria (64).

Four *P. oryctolagi* DNA samples were obtained from previous studies of one rabbit with severe combined immunodeficiency at the University of Michigan, Ann Arbor, MI, USA (65), and three immunosuppressed rabbits at the Institut Pasteur de Lille (66) and the Institut National de la Recherche Agronomique de Tours Pathologie Aviaire et Parasitologie, Tours (67), France.

Animal experimentation guidelines of the National Institutes of Health were followed in the conduct of these studies.

**Illumina sequencing.** DNA extracts for 4 *P. wakefieldiae*, 4 *Pneumocystis* sp. "*macacae*," 2 *Pneumocystis* sp. "*canis*," and 4 *P. oryctolagi* samples were subjected to whole-genome sequencing commercially



in an Illumina HiSeq platform using a 150-bp paired-end library and/or a 250-bp paired-end library. Genome assembly was performed essentially as previously described (1, 68); detailed analyses of these genomes will be published separately.

**RNA-Seq analysis of different *msg* families in *P. murina* and *P. carinii*.** The relative expression level for each gene was estimated using RNA-Seq data from three heavily infected laboratory animals each for *P. murina* and *P. carinii* as previously described (1). RNA-Seq reads from each of the three samples were aligned to the coding DNA sequences (CDSs) using bowtie (69). The alignment bam files were then used to quantify transcript abundances by RSEM (70). The relative expression level for each gene was expressed as fragments per kilobase of exon per million fragments mapped (FPKM).

***msg* sequences of *P. wakefieldiae*.** To amplify the repertoire of the classical *msg*-A1 genes in full length, the forward primer (WSG.f3) was designed from the 3' end of the previously reported UCS (within CRJE) of *P. wakefieldiae* (GenBank accession no. AF164562) (30). The reverse primer (WSG.r5) was designed from the highly conserved 3'-end coding region near the stop codon based on an alignment of more than 3,000 Illumina HiSeq reads. Primer sequences are listed in Table S2. Both primers were specific for *P. wakefieldiae*, with no cross-amplification of *P. carinii*. PCR was performed using *P. wakefieldiae* genomic DNA and the AccuPrime Pfx SuperMix kit (Thermo Fisher Scientific) with the following cycling conditions: 95°C for 5 min and then 35 cycles at 95°C for 15 s, 55°C for 30 s, and 68°C for 3 min, with a final extension at 68°C for 5 min. The PCR product was subcloned into the pCR2.1 TOPO vector by use of the TOPO TA Cloning kit (Invitrogen, Carlsbad, CA). Two clones containing the full-length *msg*-A1 gene were sequenced commercially by Sanger sequencing.

To sequence the *P. wakefieldiae* homologue of the 6-gene cluster of the *msg*-C family in *P. murina*, we first used the Illumina reads of *P. wakefieldiae* (mixed with *P. carinii* reads) to assemble the *P. wakefieldiae* homologues of PNEG\_03432 and PNEG\_03438, which are flanking the 6-gene cluster in *P. murina*. Subsequently, we designed a primer set (3432.f1 and 3438.r1) (Table S2) specific for these two genes in *P. wakefieldiae*. With these two primers, we amplified an 8-kb fragment from *P. wakefieldiae* DNA and sequenced its full length by Sanger sequencing with primer walking. From a draft *P. wakefieldiae* genome assembly, we identified members of the *msg*-A2, *msg*-A3, *msg*-D, and *msg*-E families or subfamilies based on homology to known genes in *P. murina*, *P. carinii* and *P. jirovecii* (1). Full-length *msg*-A1 genes sequences were unable to be assembled from the short HiSeq reads (16).

***msg*-A1 sequences of *P. carinii* from wild rats.** To determine whether the *msg*-A1 repertoires are identical in *P. carinii* from wild and laboratory Norway rats, we performed RFLP analysis of *P. carinii* isolates from 8 wild rats in comparison with those from 8 laboratory rats. The *msg*-A1 repertoires were amplified by PCR using primers RSG.f10 and RSG.r8 (Table S2), which are located in the highly conserved regions at the beginning and end of the *msg* coding regions, respectively, among 57 known full-length *msg*-A1 genes in *P. carinii* (1). Amplification was performed using the LongAmp Master Mix (New England Biolabs) with the following parameters: 94°C for 2 min and then 35 cycles at 94°C for 15 s, 55°C for 30 s, and 68°C for 3 min, with a final extension at 68°C for 5 min. PCR products were purified and subjected to restriction digestion with DraI (New England BioLabs) at 37°C for 2 h. The resulting digests were purified and separated in 2% E-gel containing ethidium bromide (Invitrogen, Carlsbad, CA).

The *msg*-A1 repertoire from one wild rat (no. R5), which showed a distinct RFLP pattern compared to those of laboratory rats, was chosen for sequencing after PCR amplification using the primer pair RSG.f10-RSG.r8 and the LiSpark Max SuFi PCR Master Mix kit (LifeSct LLC, Rockville, MD). The PCR product was subcloned into the pCR-XL-2 TOPO vector by use of the TOPO XL-2 Complete PCR Cloning kit (Invitrogen, Carlsbad, CA). A total of 28 clones containing the full-length *msg*-A1 gene were sequenced commercially by Sanger sequencing.

***msg* sequences of *Pneumocystis* sp. "*macacae*," *Pneumocystis* sp. "*canis*," and *P. oryctolagi*.** Illumina HiSeq reads from one *Pneumocystis* sp. "*macacae*" sample were aligned to all known full-length *msg*-A1 genes of *P. jirovecii* (1), resulting in at least 1,000 reads for the very 5' and 3' ends of the *msg*-A1 coding regions. Two primers (KSG.f3 and KSG.r2) (Table S2) were designed from highly conserved regions based on alignment of these reads. The full-length *msg*-A1 repertoire in *Pneumocystis* sp. "*macacae*" was amplified using these two primers and the LiSpark Max SuFi PCR Master Mix kit, followed by subcloning into the pCR-XL-2 TOPO vector as described above. Two clones containing the full-length *msg*-A1 gene were sequenced commercially by Sanger sequencing.

For other *msg* families and subfamilies, we identified a small number of representative genes from a partial genome assembly of *Pneumocystis* sp. "*macacae*" based on homology to known genes in *P. murina*, *P. carinii*, and *P. jirovecii* (1).

For *Pneumocystis* sp. "*canis*" and *P. oryctolagi*, a small number of genes representing each *msg* family were identified from a partial genome assemblies of *Pneumocystis* sp. "*canis*" and *P. oryctolagi*, respectively.

**UCSs of *msg*-A1 genes in *Pneumocystis* from various mammalian host species.** The UCS and its 5' untranslated region (UTR) sequences in *Pneumocystis* sp. "*macacae*," *Pneumocystis* sp. "*canis*," and *P. oryctolagi* were first obtained by assembling Illumina HiSeq reads from whole-genome sequencing as described above, followed by confirmation by PCR amplification and Sanger sequencing of genomic DNA. Based on sequence alignment of these UCSs and known UCSs of *P. murina*, *P. carinii*, *P. wakefieldiae*, and *P. jirovecii*, we designed one forward primer (5UTR) from the conserved region in the 5' UTR and one reverse primer (CRJE.r3) from the conserved region in the CRJE (Table S2). This primer set was used to amplify the UCS along with its 5' UTR in *Pneumocystis* species from other mammal species, including dogs, rabbits, chestnut white-bellied rats, Müller's giant Sunda rats, Asian house rats, and Polynesian rats (Fig. 4 and Table S1). To study the variability of UCSs and downstream *msg*-A1 coding regions in different *P. oryctolagi* isolates, PCR was performed using a pair of primers, OSG.f3 and OSG.r9, which are located

at the very 5' end of UCS and one highly conserved region near the 5' end of the *msg*-A1 coding region (Table S2). All PCR products were sequenced directly and/or after subcloning into the pCR2.1 TOPO vector as described above.

**Phylogenetic analysis.** To analyze phylogenetic relationships, deduced protein sequences were aligned using MUSCLE (71), and phylogenetic trees were constructed based on maximum likelihood (ML) using RAxML (v8.2.5) (72) with 100 bootstraps as support values. The best amino acid model was estimated using the PROTGAMMAAUTO option.

**Data availability.** Annotated genomic sequences of all new *msg* genes identified in this study are available from the BioProject database (<https://www.ncbi.nlm.nih.gov/bioproject>) under accession no. PRJNA560924. All new UCS sequences obtained in this study are available from GenBank with accession no. MN509813 to MN509830. Coding DNA sequences (CDSs) and deduced amino sequences for all *msg* genes according to the family/subfamily are available at <https://zenodo.org/record/3523554#.XjLZ7UBFyF4> (excel file for data sets 1 to 8). Hidden Markov models (HMMs) for *Msg* domains are available at <https://zenodo.org/record/3515473#.XjLaaEBFyF4>.

## SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

**TEXT S1**, DOCX file, 0.1 MB.

**FIG S1**, EPS file, 2.7 MB.

**FIG S2**, EPS file, 2.6 MB.

**FIG S3**, EPS file, 2.7 MB.

**FIG S4**, EPS file, 2.6 MB.

**FIG S5**, EPS file, 2.7 MB.

**FIG S6**, EPS file, 2.6 MB.

**FIG S7**, EPS file, 2.7 MB.

**TABLE S1**, DOCX file, 0.1 MB.

**TABLE S2**, DOCX file, 0.1 MB.

## ACKNOWLEDGMENTS

This study was funded with federal funds from the Intramural Research Program of the U.S. National Institutes of Health Clinical Center, the National Institute of Allergy and Infectious Diseases, the National Cancer Institute, National Institutes of Health, under contract number HHSN261200800001E, the National Human Genome Research Institute (grant U54HG003067 to the Broad Institute), the National Institute of Diabetes & Digestive & Kidney Diseases (grant R01DK109883 to Tulane National Primate Research Center), and the Office of Research Infrastructure Programs/OD (award P51OD011107 to CNPRC). The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

We thank Rene Costello for providing animal care, Nicolas Cere (Institut National de la Recherche Agronomique de Tours Pathologie Aviaire et Parasitologie, Tours, France) for kindly providing *P. oryctolagi* samples, and B. Scandrett, C. Roehrig, K. Konecni, and participating farmers for coordinating rat sample collection in Ontario. We also thank Phillippe Hauser (University of Lausanne, Lausanne, Switzerland) for providing information about *msg* sequences in their studies (17) and members of the Nomenclature Committee for Fungi, Scott Redhead and Konstanze Bensch, for their advice on *Pneumocystis* names.

We declare no conflicts of interests.

## REFERENCES

1. Ma L, Chen Z, Huang da W, Kutty G, Ishihara M, Wang H, Abouelleil A, Bishop L, Davey E, Deng R, Deng X, Fan L, Fantoni G, Fitzgerald M, Gogineni E, Goldberg JM, Handley G, Hu X, Huber C, Jiao X, Jones K, Levin JZ, Liu Y, Macdonald P, Melnikov A, Raley C, Sassi M, Sherman BT, Song X, Sykes S, Tran B, Walsh L, Xia Y, Yang J, Young S, Zeng Q, Zheng X, Stephens R, Nusbaum C, Birren BW, Azadi P, Lempicki RA, Cuomo CA, Kovacs JA. 2016. Genome analysis of three *Pneumocystis* species reveals adaptation mechanisms to life exclusively in mammalian hosts. *Nat Commun* 7:10740. <https://doi.org/10.1038/ncomms10740>.
2. Kutty G, Davis AS, Ferreyra GA, Qiu J, Huang da W, Sassi M, Bishop L, Handley G, Sherman B, Lempicki R, Kovacs JA. 2016. Beta-glucans are masked but contribute to pulmonary inflammation during *Pneumocystis* pneumonia. *J Infect Dis* 214:782–791. <https://doi.org/10.1093/infdis/jiw249>.
3. Gigliotti F. 1992. Host species-specific antigenic variation of a mannosylated surface glycoprotein of *Pneumocystis carinii*. *J Infect Dis* 165: 329–336. <https://doi.org/10.1093/infdis/165.2.329>.
4. Gigliotti F, Ballou LR, Hughes WT, Mosley BD. 1988. Purification and

- initial characterization of a ferret *Pneumocystis carinii* surface antigen. *J Infect Dis* 158:848–854. <https://doi.org/10.1093/infdis/158.4.848>.
5. Haidaris PJ, Wright TW, Gigliotti F, Haidaris CG. 1992. Expression and characterization of a cDNA clone encoding an immunodominant surface glycoprotein of *Pneumocystis carinii*. *J Infect Dis* 166:1113–1123. <https://doi.org/10.1093/infdis/166.5.1113>.
  6. Kovacs JA, Powell F, Edman JC, Lundgren B, Martinez A, Drew B, Angus CW. 1993. Multiple genes encode the major surface glycoprotein of *Pneumocystis carinii*. *J Biol Chem* 268:6034–6040.
  7. Tanabe K, Takasaki S, Watanabe J, Kobata A, Egawa K, Nakamura Y. 1989. Glycoproteins composed of major surface immunodeterminants of *Pneumocystis carinii*. *Infect Immun* 57:1363–1368. <https://doi.org/10.1128/IAI.57.5.1363-1368.1989>.
  8. Wright TW, Bissoondial TY, Haidaris CG, Gigliotti F, Haidaris PJ. 1995. Isoform diversity and tandem duplication of the glycoprotein A gene in ferret *Pneumocystis carinii*. *DNA Res* 2:77–88. <https://doi.org/10.1093/dnares/2.2.77>.
  9. Maddison SE, Hayes GV, Ivey MH, Tsang VC, Slemenda SB, Norman LG. 1982. Fractionation of *Pneumocystis carinii* antigens used in an enzyme-linked immunosorbent assay for antibodies and in the production of antiserum for detecting *Pneumocystis carinii* antigenemia. *J Clin Microbiol* 15:1029–1035. <https://doi.org/10.1128/JCM.15.6.1029-1035.1982>.
  10. Garbe TR, Stringer JR. 1994. Molecular characterization of clustered variants of genes encoding major surface antigens of human *Pneumocystis carinii*. *Infect Immun* 62:3092–3101. <https://doi.org/10.1128/IAI.62.8.3092-3101.1994>.
  11. Wright TW, Gigliotti F, Haidaris CG, Simpson-Haidaris PJ. 1995. Cloning and characterization of a conserved region of human and rhesus macaque *Pneumocystis carinii* gpA. *Gene* 167:185–189. [https://doi.org/10.1016/0378-1119\(95\)00704-0](https://doi.org/10.1016/0378-1119(95)00704-0).
  12. Keely SP, Stringer JR. 2009. Complexity of the MSG gene family of *Pneumocystis carinii*. *BMC Genomics* 10:367. <https://doi.org/10.1186/1471-2164-10-367>.
  13. Keely SP, Renauld H, Wakefield AE, Cushion MT, Smulian AG, Fosker N, Fraser A, Harris D, Murphy L, Price C, Quail MA, Seeger K, Sharp S, Tindal CJ, Warren T, Zuiderwijk E, Barrell BG, Stringer JR, Hall N. 2005. Gene arrays at *Pneumocystis carinii* telomeres. *Genetics* 170:1589–1600. <https://doi.org/10.1534/genetics.105.040733>.
  14. Kutty G, Maldarelli F, Achaz G, Kovacs JA. 2008. Variation in the major surface glycoprotein genes in *Pneumocystis jirovecii*. *J Infect Dis* 198:741–749. <https://doi.org/10.1086/590433>.
  15. Jiao X, Zheng X, Ma L, Kutty G, Gogineni E, Sun Q, Sherman BT, Hu X, Jones K, Raley C, Tran B, Munroe DJ, Stephens R, Liang D, Imamichi T, Kovacs JA, Lempicki RA, Huang DW. 2013. A benchmark study on error assessment and quality control of ccs reads derived from the PacBio RS. *J Data Mining Genomics Proteomics* 4:16008. <https://doi.org/10.4172/2153-0602.1000136>.
  16. Ma L, Raley C, Zheng X, Kutty G, Gogineni E, Sherman BT, Sun Q, Chen X, Skelly T, Jones K, Stephens R, Zhou B, Lau W, Johnson C, Imamichi T, Jiang M, Dewar R, Lempicki RA, Tran B, Kovacs JA, Huang da W. 2016. Distinguishing highly similar gene isoforms with a clustering-based bioinformatics analysis of PacBio single-molecule long reads. *BioData Min* 9:13. <https://doi.org/10.1186/s13040-016-0090-8>.
  17. Schmid-Siegert E, Richard S, Luraschi A, Muhlethaler K, Pagni M, Hauser PM. 2017. Mechanisms of surface antigenic variation in the human pathogenic fungus *Pneumocystis jirovecii*. *mBio* 8:e01470-17. <https://doi.org/10.1128/mBio.01470-17>.
  18. Edman JC, Hatton TW, Nam M, Turner R, Mei Q, Angus CW, Kovacs JA. 1996. A single expression site with a conserved leader sequence regulates variation of expression of the *Pneumocystis carinii* family of major surface glycoprotein genes. *DNA Cell Biol* 15:989–999. <https://doi.org/10.1089/dna.1996.15.989>.
  19. Kutty G, Ma L, Kovacs JA. 2001. Characterization of the expression site of the major surface glycoprotein of human-derived *Pneumocystis carinii*. *Mol Microbiol* 42:183–193. <https://doi.org/10.1046/j.1365-2958.2001.02620.x>.
  20. Sunkin SM, Stringer JR. 1996. Translocation of surface antigen genes to a unique telomeric expression site in *Pneumocystis carinii*. *Mol Microbiol* 19:283–295. <https://doi.org/10.1046/j.1365-2958.1996.375905.x>.
  21. Wada M, Sunkin SM, Stringer JR, Nakamura Y. 1995. Antigenic variation by positional control of major surface glycoprotein gene expression in *Pneumocystis carinii*. *J Infect Dis* 171:1563–1568. <https://doi.org/10.1093/infdis/171.6.1563>.
  22. Stringer JR. 2007. Antigenic variation in pneumocystis. *J Eukaryot Microbiol* 54:8–13. <https://doi.org/10.1111/j.1550-7408.2006.00225.x>.
  23. Huang SN, Angus CW, Turner RE, Sorial V, Kovacs JA. 1999. Identification and characterization of novel variant major surface glycoprotein gene families in rat *Pneumocystis carinii*. *J Infect Dis* 179:192–200. <https://doi.org/10.1086/314558>.
  24. Schaffzin JK, Sunkin SM, Stringer JR. 1999. A new family of *Pneumocystis carinii* genes related to those encoding the major surface glycoprotein. *Curr Genet* 35:134–143. <https://doi.org/10.1007/s002940050442>.
  25. Wells J, Gigliotti F, Simpson-Haidaris PJ, Haidaris CG. 2004. Epitope mapping of a protective monoclonal antibody against *Pneumocystis carinii* with shared reactivity to *Streptococcus pneumoniae* surface antigen PspA. *Infect Immun* 72:1548–1556. <https://doi.org/10.1128/iai.72.3.1548-1556.2004>.
  26. Ma L, Kutty G, Jia Q, Kovacs JA. 2003. Characterization of variants of the gene encoding the p55 antigen in *Pneumocystis* from rats and mice. *J Med Microbiol* 52:955–960. <https://doi.org/10.1099/jmm.0.05131-0>.
  27. Smulian AG, Stringer JR, Linke MJ, Walzer PD. 1992. Isolation and characterization of a recombinant antigen of *Pneumocystis carinii*. *Infect Immun* 60:907–915. <https://doi.org/10.1128/IAI.60.3.907-915.1992>.
  28. Smulian AG, Theus SA, Denko N, Walzer PD, Stringer JR. 1993. A 55 kDa antigen of *Pneumocystis carinii*: analysis of the cellular immune response and characterization of the gene. *Mol Microbiol* 7:745–753. <https://doi.org/10.1111/j.1365-2958.1993.tb01165.x>.
  29. Bishop LR, Davis AS, Bradshaw K, Gamez M, Cisse OH, Wang H, Ma L, Kovacs JA. 2018. Characterization of p57, a stage-specific antigen of *Pneumocystis murina*. *J Infect Dis* 218:282–290. <https://doi.org/10.1093/infdis/ijy099>.
  30. Schaffzin JK, Stringer JR. 2000. The major surface glycoprotein expression sites of two special forms of rat *Pneumocystis carinii* differ in structure. *J Infect Dis* 181:1729–1739. <https://doi.org/10.1086/315438>.
  31. Lugli EB, Allen AG, Wakefield AE. 1997. A *Pneumocystis carinii* multi-gene family with homology to subtilisin-like serine proteases. *Microbiology* 143:2223–2236. <https://doi.org/10.1099/00221287-143-7-2223>.
  32. Sunkin SM, Linke MJ, McCormack FX, Walzer PD, Stringer JR. 1998. Identification of a putative precursor to the major surface glycoprotein of *Pneumocystis carinii*. *Infect Immun* 66:741–746. <https://doi.org/10.1128/IAI.66.2.741-746.1998>.
  33. Deitsch KW, Lukehart SA, Stringer JR. 2009. Common strategies for antigenic variation by bacterial, fungal and protozoan pathogens. *Nat Rev Microbiol* 7:493–503. <https://doi.org/10.1038/nrmicro2145>.
  34. Katinka MD, Duprat S, Cornillot E, Méténier G, Thomarat F, Prensier G, Barbe V, Peyretailade E, Brottier P, Wincker P, Delbac F, El Alaoui H, Peyret P, Saurin W, Gouy M, Weissenbach J, Vivarès CP. 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* 414:450–453. <https://doi.org/10.1038/35106579>.
  35. Freitas-Junior LH, Bottius E, Pirrit LA, Deitsch KW, Scheidig C, Guinet F, Nehrbass U, Wellem TE, Scherf A. 2000. Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*. *Nature* 407:1018–1022. <https://doi.org/10.1038/35039531>.
  36. Fedorova ND, Khaldi N, Joardar VS, Maiti R, Amedeo P, Anderson MJ, Crabtree J, Silva JC, Badger JH, Albarraq A, Angiuoli S, Bussey H, Bowyer P, Cotty PJ, Dyer PS, Egan A, Galens K, Fraser-Liggett CM, Haas BJ, Inman JM, Kent R, Lemieux S, Malavazi I, Orvis J, Roemer T, Ronning CM, Sundaram JP, Sutton G, Turner G, Venter JC, White OR, Whitty BR, Youngman P, Wolfe KH, Goldman GH, Wortman JR, Jiang B, Denning DW, Nierman WC. 2008. Genomic islands in the pathogenic filamentous fungus *Aspergillus fumigatus*. *PLoS Genet* 4:e1000046. <https://doi.org/10.1371/journal.pgen.1000046>.
  37. Moran GP, Coleman DC, Sullivan DJ. 2011. Comparative genomics and the evolution of pathogenicity in human pathogenic fungi. *Eukaryot Cell* 10:34–42. <https://doi.org/10.1128/EC.00242-10>.
  38. Delaye L, Ruiz-Ruiz S, Calderon E, Tarazona S, Conesa A, Moya A. 2018. Evidence of the Red-Queen hypothesis from accelerated rates of evolution of genes involved in biotic interactions in *Pneumocystis*. *Genome Biol Evol* 10:1596–1606. <https://doi.org/10.1093/gbe/evy116>.
  39. Jarboui MA, Mseddi F, Sellami H, Sellami A, Mahfoudh N, Makni F, Makni H, Ayadi A. 2013. A comparison of capillary electrophoresis and direct sequencing in upstream conserved sequence region analysis of *Pneumocystis jirovecii* strains. *J Med Microbiol* 62:560–564. <https://doi.org/10.1099/jmm.0.045336-0>.
  40. Ma L, Kutty G, Jia Q, Imamichi H, Huang L, Atzori C, Beckers P, Groner G, Beard CB, Kovacs JA. 2002. Analysis of variation in tandem repeats in the intron of the major surface glycoprotein expression site of the human form of *Pneumocystis carinii*. *J Infect Dis* 186:1647–1654. <https://doi.org/10.1086/345721>.

41. Georgomanolis T, Sofiadis K, Papantonis A. 2016. Cutting a long intron short: recursive splicing and its implications. *Front Physiol* 7:598. <https://doi.org/10.3389/fphys.2016.00598>.
42. Sibley CR, Emmett W, Blazquez L, Faro A, Haberman N, Briese M, Tratzuni D, Ryten M, Weale ME, Hardy J, Modic M, Curk T, Wilson SW, Plagnol V, Ule J. 2015. Recursive splicing in long vertebrate genes. *Nature* 521:371–375. <https://doi.org/10.1038/nature14466>.
43. Abolins S, King EC, Lazarou L, Weldon L, Hughes L, Drescher P, Raynes JG, Hafalla JCR, Viney ME, Riley EM. 2017. The comparative immunology of wild and laboratory mice, *Mus musculus domesticus*. *Nat Commun* 8:14811. <https://doi.org/10.1038/ncomms14811>.
44. Leshner A, Li B, Whitt P, Newton N, Devalapalli AP, Shieh K, Solow JS, Parker W. 2006. Increased IL-4 production and attenuated proliferative and pro-inflammatory responses of splenocytes from wild-caught rats (*Rattus norvegicus*). *Immunol Cell Biol* 84:374–382. <https://doi.org/10.1111/j.1440-1711.2006.01440.x>.
45. Rothenburger JL, Himsworth CG, La Perle KMD, Leighton FA, Nemeth NM, Treuting PM, Jardine CM. 2019. Pathology of wild Norway rats in Vancouver, Canada. *J Vet Diagn Invest* 31:184–199. <https://doi.org/10.1177/1040638719833436>.
46. Bishop LR, Helman D, Kovacs JA. 2012. Discordant antibody and cellular responses to *Pneumocystis* major surface glycoprotein variants in mice. *BMC Immunol* 13:39. <https://doi.org/10.1186/1471-2172-13-39>.
47. Cushion MT, Stringer JR. 2010. Stealth and opportunism: alternative lifestyles of species in the fungal genus *Pneumocystis*. *Annu Rev Microbiol* 64:431–452. <https://doi.org/10.1146/annurev.micro.112408.134335>.
48. Baldwin AJ, Inoue K. 2006. The most C-terminal tri-glycine segment within the polyglycine stretch of the pea Toc75 transit peptide plays a critical role for targeting the protein to the chloroplast outer envelope membrane. *FEBS J* 273:1547–1555. <https://doi.org/10.1111/j.1742-4658.2006.05175.x>.
49. Condit CM, Meagher RB. 1986. A gene encoding a novel glycine-rich structural protein of petunia. *Nature* 323:178–181. <https://doi.org/10.1038/323178a0>.
50. Fusaro AF, Sachetto-Martins G. 2007. Blooming time for plant glycine-rich proteins. *Plant Signal Behav* 2:386–387. <https://doi.org/10.4161/psb.2.5.4262>.
51. Limper AH, Standing JE, Hoffman OA, Castro M, Neese LW. 1993. Vitro-nectin binds to *Pneumocystis carinii* and mediates organism attachment to cultured lung epithelial cells. *Infect Immun* 61:4302–4309. <https://doi.org/10.1128/IAI.61.10.4302-4309.1993>.
52. Pottratz ST, Paulsrud J, Smith JS, Martin WJ. 2nd, 1991. *Pneumocystis carinii* attachment to cultured lung cells by *Pneumocystis* gp 120, a fibronectin binding protein. *J Clin Invest* 88:403–407. <https://doi.org/10.1172/JCI115318>.
53. Vuk-Pavlovic Z, Standing JE, Crouch EC, Limper AH. 2001. Carbohydrate recognition domain of surfactant protein D mediates interactions with *Pneumocystis carinii* glycoprotein A. *Am J Respir Cell Mol Biol* 24:475–484. <https://doi.org/10.1165/ajrcmb.24.4.3504>.
54. Cushion MT, Keely SP, Stringer JR. 2004. Molecular and phenotypic description of *Pneumocystis wakefieldiae* sp. nov., a new species in rats. *Mycologia* 96:429–438. <https://doi.org/10.1080/15572536.2005.11832942>.
55. Kovacs JA, Halpern JL, Swan JC, Moss J, Parrillo JE, Masur H. 1988. Identification of antigens and antibodies specific for *Pneumocystis carinii*. *J Immunol* 140:2023–2031.
56. Lasbury ME, Lin P, Tschang D, Durant PJ, Lee CH. 2004. Effect of bronchoalveolar lavage fluid from *Pneumocystis carinii*-infected hosts on phagocytic activity of alveolar macrophages. *Infect Immun* 72:2140–2147. <https://doi.org/10.1128/iai.72.4.2140-2147.2004>.
57. Rothenburger JL, Rousseau JD, Weese JS, Jardine CM. 2018. Livestock-associated methicillin-resistant *Staphylococcus aureus* and *Clostridium difficile* in wild Norway rats (*Rattus norvegicus*) from Ontario swine farms. *Can J Vet Res* 82:66–69.
58. Latinne A, Bezé F, Delhaes L, Pottier M, Gantois N, Nguyen J, Blasdel K, Dei-Cas E, Morand S, Chabé M. 2018. Genetic diversity and evolution of *Pneumocystis* fungi infecting wild Southeast Asian murid rodents. *Parasitology* 145:885–900. <https://doi.org/10.1017/S0031182017001883>.
59. Ortiz AM, Flynn JK, DiNapoli SR, Vujkovic-Cvijin I, Starke CE, Lai SH, Long ME, Sortino O, Vinton CL, Mudd JC, Johnston L, Busman-Sahay K, Belkaid Y, Estes JD, Brechley JM. 2018. Experimental microbial dysbiosis does not promote disease progression in SIV-infected macaques. *Nat Med* 24:1313–1316. <https://doi.org/10.1038/s41591-018-0132-5>.
60. Patterson LJ, Daltabuit-Test M, Xiao P, Zhao J, Hu W, Wille-Reece U, Brocca-Cofano E, Kalyanaraman VS, Kalisz I, Whitney S, Lee EM, Pal R, Montefiori DC, Dandekar S, Seder R, Roederer M, Wiseman RW, Hirsch V, Robert-Guroff M. 2011. Rapid SIV Env-specific mucosal and serum antibody induction augments cellular immunity in protecting immunized, elite-controller macaques against high dose heterologous SIV challenge. *Virology* 411:87–102. <https://doi.org/10.1016/j.virol.2010.12.033>.
61. Pahar B, Kenway-Lynch CS, Marx P, Srivastav SK, LaBranche C, Montefiori DC, Das A. 2016. Breadth and magnitude of antigen-specific antibody responses in the control of plasma viremia in simian immunodeficiency virus infected macaques. *Virology* 13:200. <https://doi.org/10.1186/s12985-016-0652-x>.
62. Sukura A, Saari S, Jarvinen AK, Olsson M, Karkkainen M, Ilvesniemi T. 1996. *Pneumocystis carinii* pneumonia in dogs—a diagnostic challenge. *J Vet Diagn Invest* 8:124–130. <https://doi.org/10.1177/104063879600800124>.
63. Ma L, Imamichi H, Sukura A, Kovacs JA. 2001. Genetic divergence of the dihydrofolate reductase and dihydropteroate synthase genes in *Pneumocystis carinii* from 7 different host species. *J Infect Dis* 184:1358–1362. <https://doi.org/10.1086/324208>.
64. Weissenbacher-Lang C, Fuchs-Baumgartinger A, Klang A, Kneissl S, Pirker A, Shibly S, von Ritgen S, Weissenböck H, Künzel F. 2017. *Pneumocystis carinii* infection with severe pneumomediastinum and lymph node involvement in a Whippet mixed-breed dog. *J Vet Diagn Invest* 29:757–762. <https://doi.org/10.1177/1040638717710237>.
65. Song J, Wang G, Hoenerhoff MJ, Ruan J, Yang D, Zhang J, Yang J, Lester PA, Sigler R, Bradley M, Eckley S, Cornelius K, Chen K, Kolls JK, Peng L, Ma L, Chen YE, Sun F, Xu J. 2018. Bacterial and *Pneumocystis* infections in the lungs of gene-knockout rabbits with severe combined immunodeficiency. *Front Immunol* 9:429. <https://doi.org/10.3389/fimmu.2018.00429>.
66. Dei-Cas E, Chabé M, Moukhliis R, Durand-Joly I, Aliouat EM, Stringer JR, Cushion M, Noël C, de Hoog GS, Guillot J, Viscogliosi E. 2006. *Pneumocystis oryctolagi* sp. nov., an uncultured fungus causing pneumonia in rabbits at weaning: review of current knowledge, and description of a new taxon on genotypic, phylogenetic and phenotypic bases. *FEMS Microbiol Rev* 30:853–871. <https://doi.org/10.1111/j.1574-6976.2006.00037.x>.
67. Cere N, Polack B, Chanteloup NK, Coudert P. 1997. Natural transmission of *Pneumocystis carinii* in nonimmunosuppressed animals: early contagiousness of experimentally infected rabbits (*Oryctolagus cuniculus*). *J Clin Microbiol* 35:2670–2672. <https://doi.org/10.1128/JCM.35.10.2670-2672.1997>.
68. Cisse OH, Pagni M, Hauser PM. 2012. *De novo* assembly of the *Pneumocystis jirovecii* genome from a single bronchoalveolar lavage fluid specimen from a patient. *mBio* 4:e00428-12. <https://doi.org/10.1128/mBio.00428-12>.
69. Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25. <https://doi.org/10.1186/gb-2009-10-3-r25>.
70. Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323. <https://doi.org/10.1186/1471-2105-12-323>.
71. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797. <https://doi.org/10.1093/nar/gkh340>.
72. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313. <https://doi.org/10.1093/bioinformatics/btu033>.
73. Slaven BE, Meller J, Porollo A, Sesterhenn T, Smulian AG, Cushion MT. 2006. Draft assembly and annotation of the *Pneumocystis carinii* genome. *J Eukaryot Microbiol* 53 Suppl 1:S89–S91. <https://doi.org/10.1111/j.1550-7408.2006.00184.x>.